# PROBABILISTIC INDEXING

## A STATISTICAL TECHNIQUE FOR DOCUMENT IDENTIFICATION AND RETRIEVAL

M. E. MARON
J. L. KUHNS
L. C. RAY



Technical Memorandum No. 3
June 1959

DATA SYSTEMS PROJECT OFFICE
RAMO-WOOLDRIDGE a division of Thompson Ramo Wooldridge Inc.
P. O. Box 90534 Airport Station, Los Angeles 45, California

RW CMCC 79-U-2

# PROBABILISTIC INDEXING

## A STATISTICAL TECHNIQUE FOR DOCUMENT IDENTIFICATION AND RETRIEVAL

M. E. MARON
J. L. KUHNS
L. C. RAY

Technical Memorandum No. 3
June 1959

DATA SYSTEMS PROJECT OFFICE
RAMO-WOOLDRIDGE a division of Thompson Ramo Wooldridge Inc.
P. O. Box 90534 Airport Station, Los Angeles 45, California

# PROBABILISTIC INDEXING

## A Statistical Technique For Document Identification And Retrieval

(SUMMARY)

In August of last year, in an internal paper entitled "Probabilistic Indexing: A Statistical Approach to the Library Problem", one of the authors proposed a novel approach to the problem of indexing and searching for documentary data in a mechanized library. [1] By December 1958 some preliminary experiments with Probabilistic Indexing had been executed and the results were published in a Ramo-Wooldridge report entitled "Some Experiments with Probabilistic Indexing". [2] This present report describes some recent and more detailed experiments that were made to evaluate the retrieval effectiveness of this new method of literature indexing and searching.* In addition several refinements and extensions of the basic notions of Probabilistic Indexing appear in this report.**

*entitled Probabilistic Indexing.

**are presented.

[1] "Probabilistic Indexing: A Statistical Approach to the Library Problem", M. E. Maron, August 1958.

[2] "Some Experiments with Probabilistic Indexing", M. E. Maron, J. L. Kuhns and L. C. Ray, December 1958.

TABLE OF CONTENTS

LIST OF TABLES



LIST OF GRAPHS



LIST OF DIAGRAMS

# PART I.

## THE PROBLEM OF INFORMATION IDENTIFICATION AND RETRIEVAL

## (SUMMARY)

The basic function of a library computer is to accept as inputs, requests for information, and to supply as outputs, a list of those documents which are most relevant for each request. In conventional systems the information content of each document is identified by assigning to it a set of index tags and the search consists of finding those documents whose tags are logically compatible with the tags of a request. Because there is no precise relationship between the tags and the subjects that they denote, the search "strategy" which consists of matching essentially noisy tags causes the class of documents selected by a request to contain irrelevant documents and, even worse, to exclude relevant documents.

The technique of Probabilistic Indexing starts with the recognition that index terms are noisy and then introduces the mathematics of uncertainty (viz., the calculus of probability) in order to compute a probable relevance number for each document selected by a request. Probabilistic Indexing involves the assigning of weights to the index terms that are used to tag the documents of a library. These weights, in addition to statistical data concerning the library usage, are then used by the library computer so that, given a request for information, an inverse statistical inference can be made in order to derive a number (called the "relevance number") for each document, which is a measure of the relevance of the document for the requestor. The result of a search is an ordered list of those documents which satisfy the request and ranked according to their relevance number. The technique of Probabilistic Indexing is extended so that a request may be elaborated upon automatically, in the most probable direction, so as to increase the probability of selecting relevant documents, while the use of the computed relevance numbers allows irrelevant documents to be rejected.

# 1. INTRODUCTION

## 1.1 Initial Remarks About Information Retrieval

In recent years there has been increased attention given to the problem of designing, building and using an automatic library system which can accept and store large amounts of documentary data (as for example, that contained in books, journals and pamphlets of all sorts) so that the information may be retrieved rapidly upon subsequent request. The request for information might concern a single rather specific item of data or it might concern a broad class of information relevant to some desired subject matter. Regardless of the exact category that we might consider, it is quite clear that the importance of the over-all problem of information retrieval lies in the fact that information is the primary nutrient without which science, government, industry (and society itself) cannot thrive and we are set back to the extent that valuable information becomes inaccessible in our libraries.

## 1.2 Limitations and Extensions of Retrieval Systems

In what follows we have confined our attention solely to the consideration of an information retrieval system. Once the basic conceptual problems of information identification, storage and retrieval have been successfully managed one can turn attention to problems concerned with extending the range of automatic information handling. That is to say, when dealing with documentary data (expressed in ordinary language) one might like not only to store and retrieve, but, in addition, to perform the following operations on the information: automatic analysis to detect and remove redundant information, automatic abstracting of relevant information, automatic verification of information (i.e., given some items of data, decide whether or not they are inconsistent with any other data already in storage), automatic deduction (i.e., logical derivation), automatic correlation of data so as to establish trends and deviations . om trends, and so on. It appears that as a first step in the direction of general purpose information handling, as typified by the above examples, the problem of information identification and retrieval must be met and dealt with successfully.

## 1.3 A Major Difficulty

There are a number of obvious difficulties associated with the so-called "library problem" (i.e., the problem of information search and retrieval) and the one usually cited relates to the fact that documentary data are being generated at an alarming rate (the growth rate is exponential -- doubling every 12 years for some libraries) and consequently, considerations of volume alone make the problem appear frightening. However, the heart of the problem does not concern size, but, rather, it concerns meaning. That is to say, the major difficulties associated with the library problem concern the identification of information content--the problem of determining of two items of data which is "closer" in meaning to a third item--the problem of determining whether or not (or to what degree) some document is relevant to a given request, etc. In ordinary language there are no rules which prescribe how words should be selected and combined in order to express various kinds and shades of meanings. It is because ordinary language is vague and ambiguous and because there are no rules which allow us to manipulate the information on the basis of its meaning that the problem is so complex. This then is the heart of the problem and we shall have more to say about it subsequently.

## 1.4 Levels of Consideration

The problem of an automatic library can be examined at several levels ranging from the equipment frame of reference to a basic information flow perspective. There have been a number of "hardware" solutions to the problem of library size (e.g., use of microfilm, microcards, minicards, etc.) but since the major problem is logico-linguistic, we shall cast the problem on the conceptual level. Thus, we propose to make an analysis of the logic of the problem, to describe a technique for dealing with the problem, to present some logical and experimental data to support our technique and to lay aside, at least for the present, any considerations dealing with the physical implementation of the technique.

## 2. THE CONVENTIONAL APPROACH TO AN AUTOMATIC RETRIEVAL SYSTEM

### 2.1 The Role of Indexes

Because, at least for the immediate future, no machine can actually read a document and decide whether or not its subject matter relates to some given request subject, it is necessary to use some intermediate identifying tags; namely, an indexing system. An index to a document acts as a tag by means of which the information content of the document in question may be identified. The index may be a single term or a set of terms which together tag or identify the content of each document. The class of terms (whether it be a classification indexing system, coordinate indexing, etc.) which constitutes the allowable vocabulary for indexing documents in a library is the common language which bridges the gap between the information in the documents and the information requirements of the users.

### 2.2 The Assignment of Indexes

In principle, an indexer reads an incoming document and then selects one or several of the index terms from the "library vocabulary" and he coordinates the selected terms with the given document (or its accession number). Thus, the assignment of terms to each document is a go or no-go affair--for each term either it applies to the document in question or it does not. Furthermore, the process of indexing information and that of formulating a request for information are symmetrical in the sense that just as the subject content of a document is identified by coordinating to it a set of index terms, so also, the subject content of a request must be identified by coordinating to it a set of index terms. Thus, the user who has some particular information need identifies this need in terms of a library request consisting of one or several index terms or logical combinations thereof.

## 2.3 The Notion of Semantic Noise

The correspondence between the information content of a document and its set of indexes is not exact because it is extremely difficult to specify precisely the subject content of a document by means of one or several index words. If we consider the set of all index terms, on the one hand, and the class of subjects that they denote, on the other hand, then we see that there is no strict one-one correspondence between the two. It turns out that given any term there are many possible subjects that it could denote (to a greater or lesser extent) and conversely, any particular subject of knowledge (whether broad or narrow) usually can be denoted by a number of different terms. This situation may be characterized by saying that there is "semantic noise" in the index terms. Just as the correspondence between the information content of a document and its set of indexes is not exact, so also the correspondence between a user's request, as formulated in terms of one or many index words, and his real need (intention) is not exact. Thus, there is semantic noise in both the document indexes and in the requests for information.

One of the reasons that the index terms are noisy is due to the fact that the meanings of these terms are a function of their setting. That is to say, the meaning of a term in isolation is often quite different when it appears in an environment (sentence, paragraph, etc.) of other words. The position and frequency of other words help to clarify and specify the meanings of a given term. Furthermore, individual word meanings vary from person to person because, to a large degree, the meanings of the words are a matter of individual experience. This is all to say that when words are isolated and used as tags to index documents, it is difficult to pin down their meanings, and, consequently it is difficult to use them as such to accurately index documents or to accurately specify a request.

## 2.4 Conventional Stopgaps

Many workers in the field of library science have attempted to reduce the semantic noise in indexing by developing specialized indexing systems for different kinds of libraries. An indexing system tailored to a particular library would be less noisy than would be the case otherwise. (In a sense, to tailor an index system to a specific library is to apply the principle of an ideoglossary, as it is used in machine language translation, to remove semantic ambiguity.) In spite of careful work in the developing of a "best" set of tags for a particular library, the problem of semantic noise and its consequences remain, albeit to a lesser extent.

Another attempt to remove the semantic noise in request formulations has to do with the use of logical combinations of index terms. That is to say, if two or more index terms are joined conjunctively, it helps to narrow or more nearly specify a subject. On the other hand, the same set of terms connected disjunctively broadens the scope of a request.[1] Thus, using logical combinations of index terms one would hope to either avoid the retrieval of irrelevant material or avoid missing relevant material. However, although a request using a set of index terms joined conjunctively does decrease the probability of obtaining irrelevant material, it also increases the probability of missing relevant material. The converse holds for a request consisting of a disjunction of index terms.

## 2.5 The Selection Function and Some Consequences

We have said that documents are indexed by assigning one or several index terms to each, and, similarly, a library request for information is formulated by selecting one or several of those index terms which most closely denote the desired information need. Given a request, the

---

[1] Strictly speaking, the terms "intersection" and "union" should be used instead of "conjunction" and "disjunction", respectively, since we are referring to classes and not propositions.

next step is to search and select all those documents (or their accession numbers) whose sets of index terms are logically compatible with those of the request. Thus, conventional machine searching consists of matching the indexes and the requests exactly. The actual matching procedure is a go or no-go affair--a set of index terms (associated with a particular document) either satisfies a request or it does not.

The fact that conventional selecting (searching) consists in deciding whether an exact logical match exists between classes of essentially noisy tags implies that the result of a search does not provide an optimal list of documents. The fact that conventional searching consists in matching noisy tags implies that the result of a search provides documents which are irrelevant to the real needs of the requestor, and, even worse, some of the really relevant documents are not retrieved. If one broadens a request (by using more general terms) so as to reduce the probability of missing a relevant document, he increases the probability of obtaining irrelevant material. Conversely, if he narrows his request (by using rather specific terms) in order to avoid irrelevant material, he increases the probability of missing relevant information. This undesirable situation is not helped by the fact that the list of documents which results from a search appears in a random order; i.e., there is no hint given to the requestor that some of the documents that have been selected are less relevant to the request than others.

In the following section we shall present the basic notions of the technique of Probabilistic Indexing and show that this approach to the library problem improves retrieval effectiveness both by reducing the probability of obtaining irrelevant documents and by increasing the probability of selecting relevant documents. Furthermore, the technique of Probabilistic Indexing provides as the result of a search an ordered list of those documents which satisfy the request, ranked according to relevance.

## 3. THE BASIC NOTIONS OF PROBABILISTIC INDEXING

### 3.1 The Probabilistic Nature of the Problem

To say that index tags are noisy is to say that there is an uncertainty about the relationship between the terms and the subjects denoted by the terms. That is to say, given a document indexed with its assigned index term (or terms), there is only a probability that if a user is interested in the subject (or subjects) designated by the tag, he will find that the document in question is relevant. This situation is analogous to the case when a message is selected and transmitted over an electrical communication channel which is noisy, and, as a result, there is only a probability that the selected message will be received at the other end of the channel. Thus, given any arbitrary received message there is a distribution which describes the probability that it (i.e., the received message) resulted from each of the possible transmitted messages. Communication theory tells us that the ideal receiver is one that makes an inverse inference and computes, given the received message, the most probable message that was transmitted.

Again, one may consider by analogy that the documents of a library are the messages that are selected for transmission, that the indexer is the noisy channel, and that the index terms are the messages that are received after passing through the channel. By analogy, the ideal searching system is the one that makes an inverse inference and computes, given the index terms of a request, the most probable document that is relevant to the request. Given this analogy between searching a library of documents and communicating in the presence of noise we see that the real problem is to introduce the proper probabilities so that the necessary inverse statistical inference can be computed.

### 3.2 The Notion of Weighting Index Terms

We have suggested that the ideal search system is one that computes the distribution which describes the probability that a document will satisfy a requestor. This means that given a request, a class of documents is

selected (namely those whose index terms are logically compatible with the terms and logic of the request) and for each document in this class, the system will have to compute a number, called the "relevance number" which will be a measure of the expected degree of relevance of the document for the requestor. How should such a relevance number be derived? Surely, it should be a function of the probability that if a person is interested in the content of a given document then he will use the tag (or tags) associated with the document in requesting information on that subject. How to estimate this probability?

As we have stated previously, conventional indexing consists in having an indexer decide on a yes-no basis whether or not a given term applies for a particular document. Either a tag is applicable to the document or it is not--there is no middle ground. It is much more reasonable and realistic to make this judgment on a probabilistic basis; i.e., to assert that a given tag may hold with a certain probability or weight. Properly scaled this weight can be used as an estimate of the above probability; viz., the probability that if an individual desires information of the type contained in the document then he will use the tag in question in requesting that information. The details are given in Part II, 1.

Given the ability to weight index terms, one can characterize more precisely the information content of a document. The indexer may wish to assign a low weight such as 0.1 or 0.2 to a term rather than to say that the term does not hold at all for the document. Conversely, the indexer may wish to assign a weight of 0.8 or 0.9 to a term rather than to say that it definitely holds for a document. Thus, given weighted (probabilistic) indexing it is possible to more accurately characterize the information content of a document. The notion of weighting the index terms that are assigned to documents and using these weights to compute relevance numbers is basic to the technique which we call "Probabilistic Indexing".

### 3.3 The A Priori Probability Distribution

One of the major goals of the method of Probabilistic Indexing is to compute a relevance number for each document on the basis of a given request. The retrieved documents will be ordered according to their relevance numbers and hence the outcome of a search will be a list of those documents whose index terms satisfy the request; the documents will be ranked according to the probability of satisfying the request, thereby providing the user with an optimal search strategy in reading the retrieved information.

We have indicated that the relevance number of some document $D_i$ should be a function of the probability that if an individual desires information of the kind contained in $D_i$, he will use the tags associated with $D_i$ in his request for information. We have indicated further that this probability can be estimated by an indexer, and in fact, the weight of a tag (i.e., the degree with which it holds for a document) is, when properly scaled, an estimate of the above probability. (This will be discussed more completely in Part II, 1 with an explanation of how the initial estimates can be modified so as to approach the correct probability.) In addition we assert that the relevance number should also be a function of the a priori probability distribution of document usage. The a priori probability distribution of usage of documents in a sense describes the popularity of documents in a library. The justification for including the statistics on the a priori probability of a document in the computation for relevance number will be given in Part II, 1 also.

### 3.4 The Schema for Computing the Relevance Number

Although the details of the logical and mathematical justification of Probabilistic Indexing are presented in Part II we briefly summarize the theoretical motivation behind our procedure, for the reader's convenience. Given the a priori probability distribution of usage of documents and the statistical indexing information for each document, the actual search would involve an inverse probability calculation, so-called

Bayes' Theorem. This inverse probability calculation computes the probability that a document satisfies the request. The situation can be presented schematically as follows:

$P(A, D_i)$ — the a priori probability that the $i^{th}$ document will be retrieved.

$P(A, D_i, I_j)$ = the weight with which the $i^{th}$ document is indexed with the $j^{th}$ index term.

$P(A, I_j, D_i)$ = the probability that, if the $j^{th}$ index term is requested, the $i^{th}$ document will satisfy the request.

$$P(A, I_j, D_i) = \frac{P(A, D_i) \cdot P(A, D_i, I_j)}{P(A, I_j)}$$

Thus the inverse probability calculation will be made for each document which is indexed with the index word. For each of those a number will be computed which will be a function of both the degree to which the document is indexed by the given index term and also the relative frequency of usage of the document. Once these computations have been made these numbers and the associated document accession numbers will be sorted so that the document which has the highest probability of satisfying the request will appear first on the list, and that document with the lowest probability of satisfying a request will be last on the list.

### 3.5 Request Weights

Just an indexer may coordinate a weight to an index term (to indicate the degree that the tag in question holds for a given document), so also, the library system should allow a user to coordinate weights to those index terms that he uses in formulating his request for information. Just as weighted index tags allow the indexer to characterize the information content of a document more precisely, so also, weighted request tags provide additional precision to the formulation of a library request.

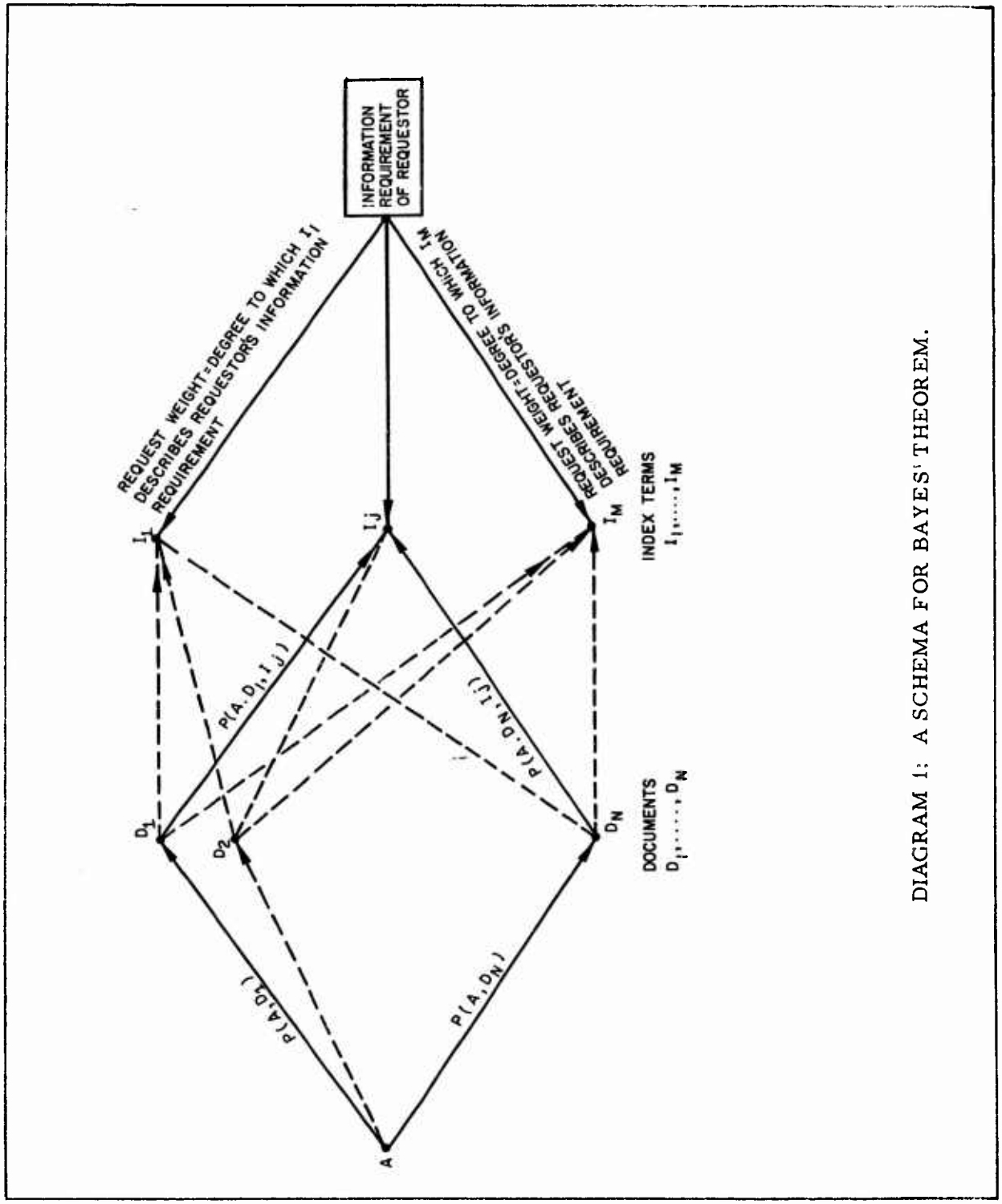



DIAGRAM 1: A SCHEMA FOR BAYES' THEOREM.

Those subjects which are most important to a user's needs will have high weights coordinated with their tags and conversely. The methods of Probabilistic Indexing provide rules describing how request weights are to be used in the extended computation for relevance number.

### 3.6 The Automatic Elaboration of a Request

Roughly speaking, we can say that the set of index terms (and their weights) identify the information content of each document with which it is coordinated and, likewise, the request formulated in terms of weighted index tags and associated logic identifies the user's information need. The next step in automatic retrieval is to match identifications in order to determine which documents are to be selected, retrieved and given to the user. Given the class of selected documents, the computation of relevance number allows a library system to rank the documents according to their probable relevance to the requestor. It is clear however, that if the initial request is inadequately (or incompletely) formulated, then the class of selected documents will not be optimal and no amount of ranking by relevance will correct this difficulty. As a remedy for this situation, Probabilistic Indexing includes methods for automatically elaborating upon any arbitrary request so as to improve its selectivity. That is to say, included among the methods of Probabilistic Indexing are mechanical rules for automatically relating index terms (and documents) so that given a request for a particular set of index terms a computer can determine what other terms are most closely related to the request and thereby automatically elaborate upon it in the most probable direction, in order to improve the selection. The rules involve the derivation of probabilistic weighting factors between index terms and a number of machine "strategies" for deciding how to go from a given request to its proper elaboration.

# PART II.

## A THEORETICAL DISCUSSION

## (SUMMARY)

The conceptual framework of the library which grows out of the basic notions of Probabilistic Indexing allows us to divide the over-all problem of information searching and retrieval into two parts. The first part relates to the problem of selecting an optimal class of documents from the entire library in order to satisfy a given request for information. We call this the selection problem. Once a class has been selected each document in it is ranked according to its probable relevance. We call this the problem of the relevance number.

Part II of this report contains the logico-mathematical analysis, explanation and (a priori) justification of the methods of Probabilistic Indexing. In particular, section 1 discusses the notion of probable relevance number and provides the detailed explication of this notion in terms of the theorem of Bayes. We discuss the meaning of index and request weights and the computational rules which allow the relevance number to be computed. Section 2 provides the discussion of how an arbitrary request may be automatically elaborated upon in its most probable direction, in order to improve the selection of documents. We describe various statistical measures for determining the "closeness" between the terms that constitute index space and indicate how these measures can be used to elaborate on a request in order to produce an optimal selection of documents.

## 1. THE DERIVATION OF THE RELEVANCE NUMBER

### 1.1 Initial Remarks

We can clarify the library problem by considering the following two fundamental questions: (1) Given the class of documents that satisfy the logic of a request which of these is most probably relevant to the requestor, next most probably relevant, etc? (2) Interpreting a request as giving clues to the real information requirements of the requestors how can the request be elaborated in order to improve the class of retrieved documents; or, more generally, how can the document selection process be improved? In this section we will discuss the first question. In particular we will establish a measure of probable relevance (the relevance number) and show how this quantity can be computed. Our methods in establishing computational procedures involve a priori considerations as well as experimental testing. The a priori considerations play their role in the choice of schemata from the theory of probability as models for our procedures and in the statistical modification of various quantities which have been estimated initially. The problem of justification of these procedures can be considered from two aspects: (1) the experimental testing; (2) a theoretical development. Since success is the only criterion upon which a retrieval system should be judged, we see that a theoretical development is unnecessary; nevertheless it is not superfluous, for it enables us to isolate assumptions and points the way for possible refinements in the procedure. The following sections (1.2 - 1.14) present a theoretical development of the computational procedure for requests formulated as Boolean functions of the index terms.

### 1.2 Notational Matters

By "P(A, B)" we mean the probability of an event of class B occurring with reference to an event of class A. We shall be interested in the following classes of events:

a. $D_i$ : obtaining the $i^{th}$ document and finding it relevant.

b. $I_j$: requesting information on the field of interest (subject, area of knowledge) designated by the $j^{th}$ index term $I_j$. (We use the same symbol for the event class and the index term, but the proper meaning will be clear in context.)

c. A: requesting information from the library.

We use "$w_{ij}$" (also "$w_i(I_j)$") to denote "the degree to which the $j^{th}$ index term applies to the $i^{th}$ document." Note that the values $w_{ij}$ define a matrix called the "probabilistic matrix", where the entry in the $i^{th}$ row and the $j^{th}$ column is the weight $w_{ij}$.

## 1.3 The A Priori Probability Distribution--A Remark about the Class A

We call "$P(A, D_i)$" the a priori probability of the document $D_i$. Although this probability arises in the inverse probability calculation, to be discussed below, we prefer to introduce it on a more intuitive level as an essential ingredient of Probabilistic Indexing. In a literature search, if two documents are indexed identically then the document with the greater a priori probability of being relevant should be read first. This is the statistical analogue of "recommendation" of texts. The calculation of $P(A, D_i)$ is obtained by the processing of library statistics. How this is done will be discussed in section 1.13, but for the present we are concerned with only one restrictive condition. This condition requires a qualification of the class A. For convenience in tabulating library statistics we will not consider A to be the class of all requests but only those that yield a document relevant to the requestor. We define the termination of such an event to be when a relevant document is obtained. Thus an event of class A will be followed by one and only one event of class $D_i$; if a request R produces two relevant documents we regard this as two instances of the request, etc. The assumption that a request will produce a relevant document we can call "the axiom of completeness of the library with respect to the index terms". In particular, this assumption allows us to normalize the a priori probability distribution. Hence we assert

$$\sum_{i=1} P(A, D_i) = 1. \qquad (1)$$

## 1.4 Statement of the Problem

Given a request R we want to rank the library documents according to their probable relevance to the requestor. A function f on a set $D_1$, $D_2$, ..., $D_n$ can be used to rank the set by means of the function values:

$$f(D_{i_1}) \geqq f(D_{i_2}) \geqq \cdots \geqq f(D_{i_n}).$$

Thus we want to look for functions which somehow measure the probable relevance of a document. If this is accomplished then a library search can be represented as a transfer function from requests (the input space) to a space of functions of the variable i(i=1, ..., n) representing the accession number of a document. That is to say, a search by Probabilistic Indexing will lead to a function; the values of this function give the probable relevances of the documents. (Those i's for which this function is not zero give the accession numbers of the documents that match the logic of the request.)

## 1.5 The Probable Relevance Function

We examine first the simplest type of request; viz., $I_j$. One measure of probable relevance is given by the function

$$P(A \cdot I_j, D_i),$$

because this is the probability that a library user, making the request $I_j$, will find the $i^{th}$ document relevant. We call this the probable relevance function. Now, keeping $I_j$ fixed, this function should certainly vary as the a priori probability $P(A, D_i)$ and also vary as $w_{ij}$. Let us assume that the probable relevance varies jointly as $P(A, D_i)$ and $w_{ij}$. We obtain then

$$P(A.I_j, D_i) = a_j \cdot P(A, D_i) \cdot w_{ij} \quad (j \text{ fixed}), \tag{2}$$

where, because we have one such equation for each j, we indicate that the constant of variation $a_j$ can itself be a function of j. Thus, given $I_j$, we rank the documents according to the quantity

$$P(A, D_i) \cdot w_{ij}.$$

Equation (2) can be regarded as the fundamental principle for Probabilistic Indexing. Subsequent experiments are to be thought of as empirical testing of this principle.

## 1.6 The Meaning of the Weights

The assumption (2) allows us to give a simple interpretation to the weights. From an inverse probability calculation we have

$$P(A.I_j, D_i) = \frac{P(A, D_i) \cdot P(A. D_i, I_j)}{P(A, I_j)} . \tag{3}$$

Comparing (2) and (3) we see that

$$w_{ij} = \left(\frac{1}{a_j}\right) \cdot \left(\frac{1}{P(A, I_j)}\right) \cdot P(A. D_i, I_j) . \tag{4}$$

Now the coefficient $a_j$ can be determined from (2). For, if we sum both sides of (2) over the subscript i and note that

$$\sum_i P(A.I_j, D_i) = 1 \tag{5}$$

by the axiom of library completeness with respect to the index terms, then we obtain

$$\frac{1}{a_j} = \sum_i P(A, D_i) \cdot w_{ij}. \tag{6}$$

Thus, we have the result

$$w_{ij} = \left(\frac{\sum_i P(A, D_i) \cdot w_{ij}}{P(A, I_j)}\right) \cdot P(A. D_i, I_j). \tag{7}$$

We call the coefficient of $P(A.D_i, I_j)$ "$\beta_j$" so that

$$w_{ij} = \beta_j \cdot P(A.D_i, I_j), \tag{8}$$

where

$$\beta_j = \frac{\sum_i P(A, D_i) \cdot w_{ij}}{P(A, I_j)} .$$

## 1.7 The Modified Weight

Our theory shows immediately the possibility of correcting, in a certain sense, the values $w_{ij}$ given by the indexer. Let us modify the weights by using the factor $\beta_j$ defined above; thus:

$$\omega_{ij} = w_{ij}/\beta_j = P(A.D_i, I_j). \tag{9}$$

Now it is true that such a modification has no effect on the document ranking for a single request $I_j$, but the possibility of modification allows us to justify our computational procedure for Boolean functions of the index terms as well as to isolate certain problems in the processing of the library statistics. In a sense, modifying the weights according to (9) is a smoothing operation, for an inspection of the numerator of $\beta_j$ shows it to be the weighted mean of the $w_{ij}$ in the $j^{th}$ column of the probabilistic matrix (with weights given by $P(A, D_i)$). Such a smoothing is necessary in making weights assigned to different index terms comparable.

## 1.8 Further Remarks on the Meaning of the Weight

Formula (8) shows the weight, which we originally interpreted as the degree to which the index term applied to the document, to be intimately related to the probability $P(A.D_i, I_j)$. This is a logical consequence of assumption (2). The statistical meaning of this probability can be clarified as follows: Suppose we presented to each member of a sampling

of potential library users the document $D_i$ and asked of them if they would have used the term $I_j$ in requesting it. The resulting relative frequency in the sample would be an estimate of $P(A.D_i, I_j)$. Now formula (8) relates the semantic measure "degree to which the index term applies to the document" and the statistical measure of how the terms will be used in retrieval requests; viz., $P(A.D_i, I_j)$. Since the statistics required are not available and certain quantities must be estimated, formula (9) tells us that the indexer would do better by estimating $P(A.D_i, I_j)$, thus bringing the coefficient $\beta_j$ as close to unity as possible.

One might raise the following question at this point: If the indexer were required to estimate $P(A.D_i, I_j)$, why not estimate $P(A.I_j, D_i)$ directly, since this is the goal of the computations? Actually, this is not quite correct. As we will show in the next section the goal of the computations is the determination of $P(A.R, D_i)$ where R is any Boolean function of the index terms. This quantity must be expressed in terms of probabilities each involving one and only one index term. The only way to do this is to transform $P(A.R, D_i)$ so that R goes into the attribute class, but then the result involves the probabilities $P(A.D_i, I_j)$. In other words, we always need the quantities $P(A.D_i, I_j)$ but hardly ever the quantities $P(A.I_j, D_i)$. The second argument in favor of the estimation of $P(A.D_i, I_j)$ over $P(A.I_j, D_i)$ appears when we consider the consistency of the comparative values. The indexer looks at each document, then runs through the various possible index terms which apply. In general $P(A.D_i, I_j)$ will vary over a much larger range than $P(A.I_j, D_i)$ as j varies and therefore it is easier psychologically for the indexer to correctly rank the values over the larger range. Furthermore, errors in the weighting (due to the estimation of $P(A.D_i, I_j)$) will have a smaller effect on the final computation than if $P(A.I_j, D_i)$ is estimated initially and $P(A.D_i, I_j)$ computed subsequently.

## 1.9 Requests as Boolean Functions

Before looking at the computational procedure for deriving the relevance number given any arbitrary request R, we must explain the meaning of the language of the request. We allow two logical operations between index terms; viz., "and" and "or". We abbreviate "$I_1$ or $I_2$" by "$I_1 \vee I_2$", "$I_1$ and $I_2$" by "$I_1 . I_2$"; the first is called a disjunctive request, the second, a conjunctive request. We now ask: If "$I_1$" and "$I_2$" are names of subjects, can "$I_1 . I_2$" and "$I_1 \vee I_2$" also be names of subjects? As a matter of fact it is convenient to answer this in the affirmative. The different interpretations of the logical combinations $I_1 . I_2, I_1 \vee I_2$, as used in request formulations are shown in Table 1.

| Request: | $I_1 . I_2$ | $I_1 \vee I_2$ |
|---|---|---|
| Logical Meaning | User requests information on the "subject" designated by $I_1 . I_2$ | User requests information on the "subject" designated by $I_1 \vee I_2$ |
| Retrieval Instruction Meaning | Search for documents indexed under $I_1$ and $I_2$ | Search for documents indexed under $I_1$ and search for documents indexed under $I_2$ |
| Class Meaning | User obtains documents indexed under both $I_1$ and $I_2$ | User obtains documents indexed under $I_1$ or $I_2$ or both |

Table 1. Interpretation of Logical Connectives

Note how the "v" inside a retrieval prescription becomes an "and" in the retrieval instructions. We can say that a disjunctive request is actually several requests but the searches are to be conducted simultaneously. The class meaning defined above has a simple geometric interpretation (a Venn diagram):

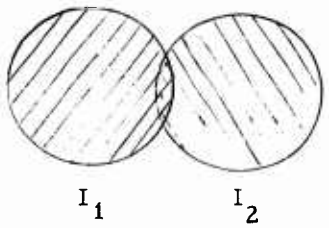



Diagram 2. Venn Diagram

The circle on the left represents those documents indexed under $I_1$, on the right those indexed under $I_2$. The overlap gives all documents obtained by requesting $I_1 . I_2$, and the entire area all those documents obtained by requesting $I_1 \vee I_2$.

## 1.10 The Extension of the Weight Function

By extending the notation for a request to include logical combinations of tags, we can consider every request R (i.e., every Boolean function of index terms), as an event class. For example $R = I_j$, $R = I_j . I_k$, $R = I_j \vee I_k$, etc. By the development in section 1.6, we see that if it is possible to compute $P(A . D_i, R)$ then we can rank documents according to probable relevance by taking the relevance number to be

$$P(A, D_i) \cdot P(A . D_i, R);$$

for, by the inverse probability calculation

$$P(A . R, D_i) = \left( \frac{1}{P(A, R)} \right) \cdot P(A, D_i) \cdot P(A . D_i, R), \qquad (10)$$

so that $P(A . R, D_i)$ is proportional to $P(A, D_i) \cdot P(A . D_i, R)$. Now we note that by (9), $P(A . D_i, R)$ is an extension of the modified weight function in the sense that:

If

$$R = I_j,$$

then,

$$\omega_{ij} = \omega_i(R) = P(A.D_i, R). \tag{11}$$

Thus the problem is to extend the function $\omega_i(I_j)$, whose values are given only for $I_1, \ldots, I_n$, to any Boolean function of these terms. We denote this extension by "$\omega_i(R)$" and we require this extension to satisfy the rules of probability since we intend for it to be an estimate of $P(A.D_i, R)$. In particular, we require:

$$0 \leqq \omega_i(R) \leqq 1, \tag{12}$$

$$\omega_i(I_1 . I_2) \leqq \omega_i(I_1), \tag{13}$$

$$\omega_i(I_1 \vee I_2) + \omega_i(I_1 . I_2) = \omega_i(I_1) + \omega_i(I_2). \tag{14}$$

We note the important fact that (14) allows us to compute the weight of a disjunction if the weight of a conjunction is known. Successive applications of (14), combined with logical transformations, allow the weight of any request to be written as additions and subtractions of weights of single terms or conjunctions. Thus the problem of the extension of the weight function is reduced to the extension to conjunctions. For these weights we also have certain restrictive conditions. If we let $p = \omega_i(I_1)$ and $q = \omega_i(I_2)$, then it can be shown that $\omega_i(I_1 . I_2)$ must be less than or equal to the minimum of the two numbers p and q and must be greater than or equal to $p + q - 1$ if this is positive, otherwise it must be greater or equal to 0. We write this condition as:

$$\max \left[0, p + q - 1\right] \leqq \omega_i(I_1 . I_2) \leqq \min \left[p, q\right]. \tag{15}$$

We have decided to take as the initial $\omega$-value of a conjunction its independence value; i.e.,

$$\omega_i(I_1 . I_2) = \omega_{i1} . \omega_{i2}. \tag{16}$$

The relevance number for a conjunction $I_1 \cdot I_2$ is then given by

$$P(A, D_i) \cdot \omega_{i1} \cdot \omega_{i2},$$

and the relevance number for a disjunction $I_1 \vee I_2$ becomes by (14)

$$P(A, D_i) \cdot \left[ \omega_{i1} + \omega_{i2} - \omega_{i1} \cdot \omega_{i2} \right] .$$

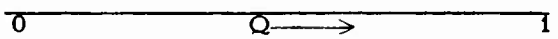
Several remarks need to be made about use of the independence value. Note that we do not say that the tags are independent--in fact they are not--but the word 'estimate' is useful to avoid making a false assumption. First, we estimate $\omega_i(I_i \cdot I_2)$ by $\omega_{i1} \cdot \omega_{i2}$. Second, we use the independence value relative to the class $D_i$, that is, we take

$$P(A \cdot D_i \cdot I_1, I_2) = P(A \cdot D_i, I_2), \tag{17}$$

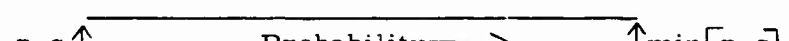
but not

$$P(A \cdot I_1, I_2) = P(A, I_2). \tag{18}$$

We believe the former estimate is more accurate than the latter. In Part II, 2.6 we discuss a coefficient of association between index terms. This coefficient which we call Q lies in the interval $\left[ -1, 1 \right]$ with Q = 0 being the point of independence. The joint occurrence of two events will have a probability in excess of its independence value only if the corresponding value of Q is positive. We have two intervals to schematize this situation: (p and q are the probabilities of the separate events and Q their coefficient of association)




Interval of Positive Association




Corresponding Interval of Probability Value

An investigation of the statistical correlation between tags via the computation of Q and then a subsequent study of which pairs of tags were used in requesting shows that Q had positive values for almost all of these pairs. This indicated that computations were called for with estimates of $\omega_i(I_1 . I_2)$ taken at the upper end of the scale; i.e., where

$$\omega_i(I_1 . I_2) = \min \left[ \omega_{i1}, \omega_{i2} \right] . \tag{19}$$

The results were not as successful as using the independence value. A possible explanation lies in noting that independence is a three term relation as formulas (17) and (18) show. It could well be that the probability value for tags $I_1$ and $I_2$ relative to the reference class A lies closer to the maximum value (min $\left[ p, q \right]$ ); while the probability value for $I_1$ and $I_2$ relative to A. $D_i$ lies closer to its independence value. For our initial estimates we assume this to be the case.

## 1.11 Estimation and Correction

We have given a formal clarification of the notions behind Probabilistic Indexing. We see that the computation of a relevance number requires for the single request the quantities $P(A, D_i)$, $w_{ij}$; for Boolean functions the quantities $P(A, D_i)$, $\omega_{ij}$, and the $\omega$-values for conjunctions. The next problem is to obtain these quantities and this, in turn, involves two problems; viz.,

(a) the initial estimation; i.e., the estimates before library statistics are obtained;

(b) the correction of the initial estimates as library statistics are accumulated and subsequent periodic revisions via a feedback computation.

These two problems are called the "problem of estimation" and the "problem of correction". Note that a solution of the second problem must be qualified by two requirements: (1) the effects of the initial estimates must die out as library statistics are accumulated, (2) the solution must not involve an impractical amount of computation.

## 1.12 The Problem of Estimation

Consider first the estimation of $P(A, D_i)$. For real libraries where no statistics are available we are confronted with this problem (as for example, in the cases when the library has not yet been used or when new documents are added). One possibility for setting the library system in motion is to take all initial values equal; i.e.,

$$P(A, D_1) = P(A, D_2) = \ldots .$$

Alternately we can construct a more realistic distribution ("more realistic" because our method leads to a distribution which corresponds more nearly to the actual distribution for some large libraries; viz., a non-linear (hyperbolic) distribution). The considerations used in such a simulation are: 1) that a correlation exists between the probability that a given tag will be used in indexing and the probability that it will be used in requesting; 2) if document $D_1$ has a higher a priori probability than document $D_2$ then $D_1$ probably has tags that $D_2$ does not have and that are used more frequently in requesting. Thus $P(A, D_i)$ should depend on the extent that $D_i$ covers the subjects designated by the library tags and also on the scope of its individual tags; i.e., on the relative frequency with which the tags are used in indexing. We therefore take as initial value of $P(A, D_i)$:

$$\alpha \sum_{j=1} N_j \cdot w_{ij},$$

where $N_j$ is the number of documents to which the $j^{th}$ index term is applied with non-zero weight, $w_{ij}$ is the weight with which the $j^{th}$ index term applies to the $i^{th}$ document, and $\alpha$ is the normalization factor (i.e., the value that gives

$$\sum_{i=1} P(A, D_i) = 1.)$$

By an argument similar to the above we find that

$$\sum_i P(A, D_i) \cdot w_{ij}$$

is a plausible estimate of $P(A, I_j)$. This estimate also has the virtue of forcing the initial value of the factor $\beta_j$ in (8), section 1.6, to equal unity. Thus, initially,

$$\omega_{ij} = w_{ij}. \tag{20}$$

The only problem that remains is the estimate of the $\omega$-values for conjunctions. We have decided to choose the value

$$\omega_i(I_1) \cdot \omega_i(I_2)$$

for $\omega_i(I_1 . I_2)$, for reasons described in section 1.10.

## 1.13 The Problem of Correction

Consider now the correction of $P(A, D_i)$. Let $P_o$ be its initial estimate. (The subscript "i" will be fixed throughout the discussion.) After n uses of the library, let $\overline{n}$ be the number of times that the $i^{th}$ document has been used. The empirical estimate of $P(A, D_i)$ is therefore $\overline{n}/n$. We want to combine this with $P_o$ in some way. Let us do so by the following device: Annex to the sequence of events of class A (i.e., $A_1$, $A_2$, ..., $A_n$) a fictitious initial sequence of length $n_o$ and suppose that this initial sequence has given the relative frequency $P_o$, then the total sequence of length $n_o + n$ will give the relative frequency

$$P_n = \frac{\overline{n} + n_o \cdot P_o}{n + n_o} . \tag{21}$$

Thus by a suitable choice of $n_o$ we can control the effect of $P_o$ on the $n^{th}$ estimate, $P_n$. For example, if $n_o = 0$ then $P_o$ has no effect on the computation ($P_n = \overline{n}/n$); while if $n_o = \infty$ then $P_o$ has its maximum effect ($P_n = P_o$).

Formula (21) gives a satisfactory estimate of $P(A, D_i)$ in a stationary system; i.e., when the reference class A does not vary with time so that

$$P(A, D_i) = \lim_{n \to \infty} P^n \tag{22}$$

holds. But in reality this will not be the case since the popularity of documents will vary with time and therefore we must look for a procedure in which the most recent statistics have the most important influence on an estimation of $P(A, D_i)$. We present a method that takes this consideration into account as well as being suitable for machine computation.

First, it will be convenient to

(a) compute $P_n$ periodically; i.e., after sequential blocks of fixed size, say m;

(b) store only statistics on the block presently occurring and the previous estimate of $P(A, D_i)$.

We propose the following computing schema: let $\overline{m}_k$ be the number of times that the $i^{th}$ document was used in the $k^{th}$ block of length m, then take as the first estimate of $P(A, D_i)$

$$P^{(1)} = \frac{\overline{m}_1 + n_o \cdot P_o}{m + n_o} \tag{23}$$

and as the $k^{th}$ estimate

$$P^{(k)} = \frac{\overline{m}_k + n_o \cdot P^{(k-1)}}{m + n_o} \tag{24}$$

Let us see how the "block" relative frequencies are involved in $P^{(k)}$. Let "A" stand for "$n_o / (m + n_o)$". Then it can be shown that

$$P^{(k)} = \left(\frac{m}{n_o}\right)\left[A\left(\frac{\overline{m}_k}{m}\right) + A^2\left(\frac{\overline{m}_{k-1}}{m}\right) + \ldots + A^k\left(\frac{\overline{m}_1}{m}\right) + A^k\left(\frac{n_o \cdot P_o}{m}\right)\right]. \quad (25)$$

Thus $P^{(k)}$ is, what is called, a convex linear combination of the relative frequencies

$$\overline{m}_k/m,\ \overline{m}_{k-1}/m,\ \ldots,\ \overline{m}_1/m,\ P_o$$

with weights

$$\left(\frac{m}{n_o}\right) A, \left(\frac{m}{n_o}\right) A^2, \ldots, \left(\frac{m}{n_o}\right) A^k, A^k .$$

That the linear combination is convex, i.e., that the sum of the weights is one, is seen from the fact that

$$\left(\frac{m}{n_o}\right) \sum_{j=1}^{k} A^j = 1 - A^k. \quad (26)$$

We see that the sequence of weights diminishes so that the more recent relative frequencies more strongly influence the value of $P^{(k)}$. We also see that

$$\lim_{k \to \infty} A^k = O \quad (27)$$

so that the effects of the earlier statistics die out as k increases. As an example we note the special case where $n_o = m$; then $A = 1/2$ and we have the weights:

$$1/2,\ 1/4,\ \ldots,\ 1/2^k,\ 1/2^k.$$

As a final remark on the computation of $P(A, D_i)$ we emphasize that caution is required in the use of our computing schema as given by (23) and (24). The possibility exists that the instances of use of the

library may attain a large number (relative to the number of documents) during a period of time when conditions are sufficiently stationary to enable us to say that (22) holds. In this case it can be shown that the recursive procedure is assured of working only if we take m so large that each value $P^{(k)}$ closely approximates $P(A, D_i)$. But in that case we would not like to be committed to $P_o$ while the first sequential block is occurring. Thus it is suggested that, for libraries of this type, we revert to some form of (21) with computations performed at shorter intervals.

The correction of $P(A, I_j)$ follows lines similar to the foregoing discussion of $P(A, D_i)$. There is, however, a rather subtle question involved in the processing of the relative frequency data. That is to say, all the probabilities are determined if we know the probabilities of conjunctions; but many requests will be given as disjunctions and thus if a library user requests information on the subject designated by $I_1 \vee I_2$ then this should affect the relative frequency of the requests $I_1$ and the requests $I_2$. A counting procedure for distinctive requests must therefore be established. The best possibility seems to be to give partial "credit" to each disjunct in an instance of a disjunction. The best way to do this is still open.

The next item is the modification of the $\omega$-values. In principle, if perfect accuracy were required, we would need the determination of $P(A, R)$ where R is any conjunctive request; for, if $\omega_i^{(k)}(R)$ is the $k^{th}$ estimate of $\omega_i(R)$ and $P^{(k)}(A, R)$ is the $k^{th}$ estimate of $P(A, R)$ then

$$\omega_i^{(k+1)}(R) = \omega^{(k)}(R) \cdot P^{(k)}(A, R) \Big/ \sum_i P^{(k)}(A, D_i) \cdot \omega_i^{(k)}(R) \qquad (28)$$

The $\omega$-values for any request would then be obtained by the extension of formula (14) mentioned in section 1.10. However, because even for a small number of index terms, the number of possible conjunctions is enormous and therefore, practical considerations would probably limit the application of formula (28). Suppose therefore, we

settle for accuracy for only the ω-values for single terms. This should be sufficient for the following reasons: (1) if we use some reasonable computational procedure--perhaps even incorporating coefficient of association data--we should obtain a sufficiently accurate relevance number to both order the documents and to execute the search strategy program of Part II, section 2, (2) the ω-value and the number of documents retrieved rapidly diminishes as the number of conjunctive terms increases--in either case accuracy in the relevance number is not required.

### 1.14 Weighted Requests

The request language is still rather limited even though we allow all combinations of index terms by means of the connectives "and" and "or". However, when we consider the retrieval instruction meaning of the request $I_1 \vee I_2$ (Table 1, p. 21) an obvious extension of the request language presents itself. That is to say, when the requestor asks that two simultaneous searches be made, one under $I_1$ and one under $I_2$, let him now indicate which search he regards as more important. To incorporate such information into a computational procedure we allow him to give this comparative data in the form of numerical "request weights". We will use the expression:

$$(\alpha)I_1 \vee (\beta)I_2,$$

where α and β are the request weights, to represent this new type of request. More generally, we note that conjunctions will occur in place of $I_1$ and $I_2$ in this expression, each conjunction prescribing a search and having an assigned weight. We can conceive of the weights, then, as indicating the degree to which the conjunction describes the information requirement of the requestor. This is suggested when we go from the retrieval instruction meaning of a request to the logical meaning (Table 1, p. 21). The highest level of specificity that the requestor can attain is by means of conjunctions. The conjunctions are (artificial) names of sub-subjects and since the requestor is uncertain about his

use of tags in forming these names, he will try to avoid possible loss of relevant information by using disjunctions. Thus when we permit him to weight each of these names; i.e., conjunctions, we can treat the weight as being an indication of either the degree of the requestor's interest in the subject designated by the conjunctive set of terms or how closely it matches his information requirements. For example, if the conjunctions reduce to single terms, then the expression

$$(.7)I_1 \vee (.3)I_2$$

means the requestor is interested in $I_1$ to the degree 0.7 and in $I_2$ to the degree 0.3.

Given this interpretation for the notion of request weights we must now provide a set of rules for computing the relevance number of weighted requests. This means that we need to evaluate

$$\omega_i\left[(\alpha)I_1 \vee (\beta)I_2\right].$$

Two methods appear to be reasonable:

(1) $\omega_i\left[(\alpha)I_1 \vee (\beta)I_2\right] = (\alpha)\omega_i(I_1) + (\beta)\omega_i(I_2) - (\alpha\beta)\omega_i(I_1 \cdot I_2)$

(2) $\omega_i\left[(\alpha)I \vee (\beta)I_2\right] = (\alpha)\omega_i(I_1) + (\beta)\omega_i(I_2) - \min\left[\alpha, \beta\right] \cdot \omega_i(I_1 \cdot I_2)$

Method (1) has the advantage of computational simplicity as well as a certain appeal in being a direct modification of the weights in the probabilistic matrix through multiplication of the request weight with the corresponding tag weight (recall that initially we take $\omega_i(I_1 \cdot I_2)$ equal to $\omega_i(I_1) \cdot \omega_i(I_2)$). Method (2) has the virtue of giving:

$$\omega_i\left[(\alpha)I_1 \vee (\alpha)I_2\right] = (\alpha)\omega_i(I_1 \vee I_2).$$

thus reducing to the case of the unweighted request when $\alpha = \beta$ (up to multiplication by a constant). In our work on the weighted request we have used formula (1).

To complete the discussion we look at further possibilities of generalization. The most obvious extension is the assignment of request weights, not just to conjunctions, but to each index term. We could explicate the meaning of

$$(a)I_1$$

as a "quantitative assertion" of $I_1$ (analogous to the "quantitative negation" in Probability Logic). The number a would be the degree of assertion given to the tag $I_1$. In this interpretation we have:

$$(0)\ I_1 = \text{not } I_1,$$
$$(1)\ I_1 = I_1.$$

The computational procedure which would be best to use here is still open.

Another possibility is a statistical or probabilistic explication of request weights; i.e., an explication of the type as given for $w_{ij}$. There, the logic was to go from a comparative notion to a quantitative notion, then to explicate the latter. We have a similar explicative problem with the request weights but no simple solution presents itself.

## 2. THE AUTOMATIC ELABORATION OF THE SELECTION PROCESS

### 2.1 Initial Remarks

The technique of Probabilistic Indexing, as we have seen, allows a computing machine, given a request for information, to make a statistical inference and derive a relevance number for each document. The result of a search is an ordered list of those documents which satisfy the request, ranked according to their probable relevance. We would prefer to have a technique which not only decides of a given class of documents, which among them is most probably relevant, next most probably relevant, etc., but which also decides whether the class itself of retrieved documents is adequate (at least in the sense of determining whether or not it excludes some documents which are relevant to the user's information needs). That is to say, if we consider the request as a clue which the user gives to the library to indicate the nature of his information needs, then we should raise the following question: Given a clue, how may it be used by the library system to generate a best class of documents (to be ranked subsequently by their relevance numbers)? Thus given the clue, how can we elaborate upon it automatically in order to produce a best class of retrieved documents? Let us turn our attention to this problem.

### 2.2 Search Strategies and the Notion of Distance

A library request (a clue) is a Boolean function whose variables are index terms, which, in turn, selects a class of documents via a logical match. That is to say, all of those documents whose index terms are logically compatible with the logic and the tags of a request R constitute the class of retrieved documents C. Our goal is to extend the class C in the most probable "direction" and this can be done in two ways. One method involves the transforming of R into R' where R' in turn will select a class of documents C', which is larger than C and contains more relevant documents. A second method does not modify R but, rather, it uses the class C to define a new class C''. A set of rules which prescribe how to go from a given request R to a class of

retrieved documents is called a strategy.[1] A strategy, in turn, involves the use of several different techniques for measuring the "distances"[2] between index terms and between documents. Before proceeding, let us introduce some further notations to make more precise what we have been saying.

We understand by "basic selection process" the rule which uses the request to select the class of documents whose tags are logically compatible with the logic and tags of the request, and we denote this basic selection process by the functional notation "f". Thus f is the transfer function from inputs (requests) to output (class of retrieved documents) and we write

$$f(R) = C \tag{1}$$

where, again, R is the request and C is the class of retrieved documents. The problem is to enlarge C so as to increase the probability that it will contain relevant documents and to decrease the probability that it will contain irrelevant documents. This can be done in the following way: Suppose R' is a request similar in meaning to R, then we can take as a possible modification of f, say f',

$$f'(R) = f(R) \vee f(R') = C \vee C'. \tag{2}$$

(As before "v" designates class union.)

---

[1] We mean here search strategy from the viewpoint of the library computer. The requestor also has a search strategy which is given by the relevance number of the documents that he is given.

[2] We use "distance" in an informal sense; i.e., it may not satisfy all the axioms of distance (e.g., the triangle inequality and symmetry). The reason for this is that it is frequently necessary to preserve natural logical structure and forego artificial metric structure. Indeed, in one case we violate positiveness of distance functions.

This modification can be made precise if we are able to invent a metric or "distance" function on the request space to measure dissimilarity in meaning. Since we are not sure what meaning is, much less being able to assign a numerical quantity to it, this is rather difficult; but we shall show later that statistics can provide such measures. For the present, suppose we actually do have such a metric, then we can generate a modified selection function f' by defining f'(R) to be the union of all classes f(R') where the distance between R and R' is less than some specified number, say $\epsilon$. Symbolically this written

$$f'(R) = \bigcup_{[d(R,R')<\epsilon]} f(R'). \tag{3}$$

Analogously, if we have a "distance" function in the document space which gives "nearness" as a numerical measure of similarity of information content, then a completely different modification f'' of f arises via

$$f''(R) = C'' \tag{4}$$

where C'' consists of all documents whose distance from C = f(R) is less than $\epsilon$. (We remark in passing that "distance" notions seem to present a surprisingly fruitful approach to the library problem; e.g., the relevance number itself can be thought of as given the nearness between documents and requests.)

Thus, we see that a machine strategy can elaborate upon the basic selection process in order to improve the search in one of two different ways. The first is to establish a metric for distance in request space so as to formulate R', given R. The other way is to use the class of documents C, obtained by the initial request R, to define a new class C''. Both of these methods are discussed below.

## 2.3 The Notion of Index Space

Geometrically speaking, one may think of the set of n index terms which constitute the library catalogue "vocabulary" as points in an

n-dimensional space. The points in this space are not located at random, but rather, they have definite relationships with respect to one another depending on the meanings of the terms. For example, the term "logic" would be much closer to "mathematics" than to "music". One always finds when looking up index terms in the catalogue of a conventional library, other terms listed under "see" and "see also". This cross-indexing ("see/see also") aspect of a library indicates some of the relationships that index terms have for one another; i.e., it indicates some of the relationships between points in index space.

The "distances" between index terms can be made explicit by formulating probabilistic weighting factors between them. Once numerical weighting factors are coordinated with the distances the cross-indexing aspect of a library can be mechanized so that given a request involving one (or many) index terms, a machine could compute other terms for which searches should be made. That is to say, a request places one at a point, or several points, in index space and once the distances between points are arithmetized, a machine could determine which other points to go to in order to improve the request. Thus, the elaboration of a request on the basis of a probabilistic "association of ideas" could be executed automatically.

## 2.4 Automatically Groping in Index Space

There are at least two different kinds of relationships that can exist between the points in index space; viz., semantical relationships and statistical relationships. The most elementary semantical relationship is that of synonymity, but in addition to synonymity there are other semantical relationships such as "partially implied by" and "partially implies". Such relationships between terms are based strictly on the meanings of the terms in question--hence, the word "semantical". Another class of relationships are statistical; i.e., those based on the relative frequency of occurrence of terms used as indexes. The distinction between semantical and statistical relationships may be clarified as follows: Whereas the semantical relationships are based solely

on the meanings of the terms and hence independent of the "facts" described by those words, the statistical relationships between terms are based solely on the relative frequency with which they appear and hence are based on the nature of the facts described by the documents. Thus, although there is nothing about the meaning of the term "logic" which implies "switching theory", the nature of the facts (viz., that truth-functional logic is widely used for the analysis and synthesis of switching circuits) "causes" a statistical relationship. (Another example might concern the terms "information theory" and "Shannon"--assuming, of course, that proper names are used as index terms.)

Once the various "connections" between the points of index space have been established rules must be formulated which describe how one should move in the maze of connected points. We call such rules "heuristics". They are general guides for groping in the "maze" in the attempt to create an optimal output list of documents for any arbitrary request. The heuristics would enable a machine to decide, for a given set of request terms, which index terms to "see" and "see also", and how deep this search should be and when to stop, etc. Generally speaking, the heuristics would decide which index terms to look at next, on the basis of the semantical and statistical connections between terms, and the heuristics would decide when to stop looking, on the basis of the number of documents that would be retrieved and the relevance numbers of those documents. (Remember that each point in index space defines a class of documents; viz., all of those documents which have been assigned the index term in question with a non-zero weight.) Given this understanding of heuristics, we see that an over-all search strategy is made up of components some of which are heuristics; i.e., the sequence of devices, rules, heuristics, etc., which lead from inputs (requests) to outputs (classes of retrieved documents) is the strategy.

## 2.5 Some Elementary Heuristics

In order to clarify the notion of developing heuristics which would determine how a computer should "grope" in index space, consider the following example. Assume that we compute the frequency, $N(I_j)$, with which each term is used to tag a document, and also that we compute the frequency, $N(I_j, I_k)$, with which pairs of terms are assigned to documents. We can then compute the conditional probability $P(I_j, I_k)$ that if a term $I_j$ is assigned to a document, then $I_k$ also will be assigned:

$$P(I_j, I_k) = \frac{N(I_j, I_k)}{N(I_j)} . \tag{5}$$

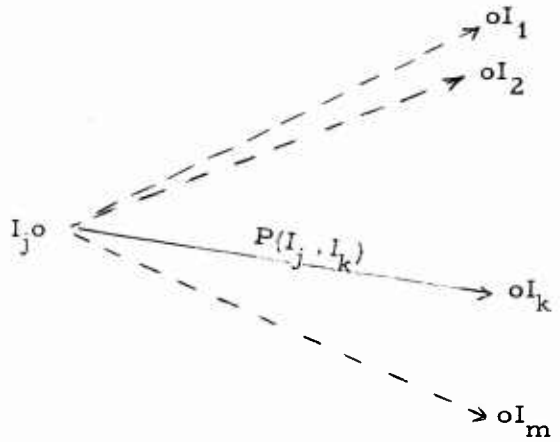
We do this for all pairs $I_j, I_k$.

Assume now that $I_j'$ is the index term which has the highest conditional probability given $I_j$; i.e., $I_j'$ is the index term for which $P(I_j, I_k)$ is a maximum. Then given a request, $R = I_j$, for all documents tagged with $I_j$, we form a new request, $R' = I_j \vee I_j'$, which searches for all documents tagged with either $I_j$ or $I_j'$. Thus, the rule is now to consider R' instead of R.

This heuristic tells us which tags are closest (in one sense) to given ones, but we still have no measure of the "closeness" (hereafter written without quotes) and such a measure is needed as a part of the associated computation rule. That is to say, we elaborate upon R and obtain R' by searching for documents indexed under tags closely related to those in the original request, but, clearly, the relevance numbers that we derive for these "additional" documents should be weighted down somewhat in order to indicate that they were obtained only from tags which are close to those in the original request. We measure the closeness as follows: Let $p_j = P(I_j, I_j')$ and normalize $p_j$ over the set of tags used in the request so that

$$\bar{p}_j = \frac{p_j}{\sum p_j} . \tag{6}$$

Now, instead of using $w_i(I_j')$ (the weight assigned to $I_j'$ for the $i^{th}$ document) in the search computation, we replace it by $\overline{p}_j \cdot w_i(I_j')$. The extended search that we have just described is an elementary form of only one of a class of possible heuristics based on the statistical relationships between tags.



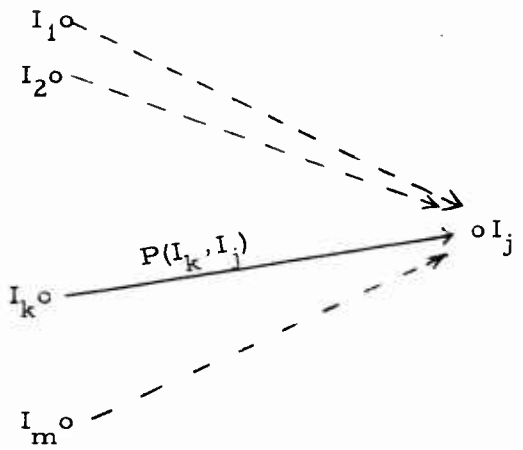

($I_j$ implies $I_k$ to the greatest degree)

A second elementary heuristic which looks even more promising is called the "inverse conditional" search and it measures closeness of tags to $I_j$ in terms of the conditional probability from $I_k$ to $I_j$ (instead of conversely as with the heuristic described above). That is to say we compute that $P(I_k, I_j)$ which is maximum as $I_k$ varies and this provides the tag which most strongly implies the given tag $I_j$. Thus, instead of asking for that tag which is most strongly implied (statistically) by an arbitrary tag in the request, we ask for the tag which most strongly <u>implies</u> (statistically) the given tag. Using this method to determine the closeness of tags we establish a measure for the closeness by normalizing the probability as before. That is,

$$p_j = P('I_j, I_j)$$

$$\overline{p}_j = \frac{p_j}{\sum p_j}$$

and, again, the corresponding computation rule is now $\overline{p}_j \cdot w_i('I_j)$, where $'I_j$ is the $I_k$ which makes $P(I_k, I_j)$ a maximum for a given $I_j$.




($I_j$ implied by $I_k$ to the greatest degree)

## 2.6 A More Sophisticated Heuristic

We have just discussed two possible measures of closeness; viz., the conditional probability $P(I_j, I_k)$, and the inverse condition probability $P(I_k, I_j)$. Now we consider a third statistical measure which appears the most promising of the three. This is one of several possible coefficients of association between predicates[1]. The particular coefficient

[1] G. U. Yule, "On Measuring Association Between Attributes", Journal of the Royal Statistical Society, Vol LXXV, 1912, pp 579-642.

we have chosen arises in the following way. Consider the tags $I_j$ and $I_k$ and partition the library by four classifications; viz., documents indexed under both $I_j$ and $I_k$, those indexed under $I_j$ but not $I_k$, those indexed under $I_k$ but not $I_j$, and those not indexed under either. Letting '$\overline{I}_j$' denote the complement of the class $I_j$, etc., these four classes are given by $I_j . I_k$, $I_j . \overline{I}_k$, $\overline{I}_j . I_k$, $\overline{I}_j . \overline{I}_k$, respectively. The classification and the number of documents is shown most conveniently in a table:

| | $I_k$ | $\overline{I}_k$ | |
|---|---|---|---|
| $I_j$ | $x = N(I_j . I_k)$ | $u = N(I_j . \overline{I}_k)$ | $N(I_j)$ |
| $\overline{I}_j$ | $v = N(\overline{I}_j . I_k)$ | $y = N(\overline{I}_j . \overline{I}_k)$ | $N(\overline{I}_j)$ |
| | $N(I_k)$ | $N(\overline{I}_k)$ | n |

We have adjoined to the table the row and column sums and n (the total number of documents).

Now, using the notation of section 2.5, we say that $I_j$ is statistically independent of $I_k$ if

$$P(I_j, I_k) = P(I_k). \tag{7}$$

This can be shown to be equivalent to:

$$P(I_j, I_k) = P(I_j) \cdot P(I_k); \tag{8}$$

so that rewriting in terms of frequencies we have an additional equivalence:

$$N(I_j . I_k) = N(I_j) \cdot N(I_k)/n. \tag{9}$$

We can infer also that the following are equivalent:

(a) $I_j$ is statistically independent of $I_k$;

(b) $I_k$ is statistically independent of $I_j$;

(c) $\overline{I}_j$ is statistically independent of $I_k$;

(d) $\overline{I}_k$ is statistically independent of $I_j$;

(e) $\overline{I}_j$ is statistically independent of $\overline{I}_k$.

For any pair $I_j, I_k$ (9) suggests that we look at the excess of $N(I_j . I_k)$ over its independence value; i.e., the quantity

$$\delta(I_j, I_k) = N(I_j . I_k) - N(I_j) \cdot N(I_k)/n. \tag{10}$$

It can be shown that this function $\delta$ has the property

$$\delta(I_j, I_k) = \delta(\overline{I}_j, \overline{I}_k) = -\delta(\overline{I}_j, I_k) = -\delta(I_j, \overline{I}_k), \tag{11}$$

and thus $\delta$ is associated with the differences over independence values in all four classifications.

Having discussed independence let us now consider what properties would be suitable for a coefficient of association between $I_j$ and $I_k$. We call this coefficient "$Q(I_j, I_k)$". (1) $Q(I_j, I_k)$ should be zero when $\delta(I_j, I_k) = 0$ and, moreover, $Q(I_j, I_k)$ should vary as $\delta(I_j, I_k)$ for fixed n and fixed row and column totals; (2) the maximum of $Q(I_j, I_k)$ should occur when $I_j$ is contained in $I_k$ ($u = 0$), or $I_k$ is in contained $I_j$ ($v = 0$), or $I_j$ and $I_k$ give the same class ($u = v = 0$); (3) the minimum of $Q(I_j, I_k)$ should occur when $I_k$ is contained in $\overline{I}_j$ ($x = 0$), or $\overline{I}_j$ is contained in $I_k$ ($y = 0$), or $I_j$ is the complement of $I_k$ ($x = y = 0$); (4) it should have a

have a simple range of values, say from -1 to 1. A coefficient[1] that has all of these properties is:

$$Q(I_j, I_k) = (xy - uv)/(xy + uv). \qquad (12)$$

(The intimate connection with $\delta$ is indicated by the fact that the numerator of Q is $n\delta$.)

The generation of a heuristic now proceeds by the plan of section 2.5. Given $R = I_j$ we select the term $I_k$ (different from $I_j$) with the maximum coefficient $Q(I_j, I_k)$. This value will be between 0 and 1 or no term will be selected. Then R is extended to

$$R = I_j \vee I_k$$

and in the search computation we multiply the weight $w_i(I_k)$ by $Q(I_j, I_k)$.

## 2.7 Heuristics in the Document Space

It seems that the modification of the selection process by means of a concept of closeness, or distance, in the request space holds the best promise for the generation of satisfactory heuristics. However, so that no possibility is overlooked, we now examine other notions of distance. Turning our attention to the modification by distance notions in the document space (equation 4, section 2.2) we see that the procedure is to go from the given request R to C, the class of documents retrieved by the basic selection process f. We then obtain C'' by applying the distance function in the document space to C. In a sense then, this two step procedure uses C as a representation of R

---

[1] The coefficient recommended by Yule loc. cit. is not Q, but

$$Z = (\sqrt{xy} - \sqrt{uv})/(\sqrt{xy} + \sqrt{uv}).$$

The range of variation of both Q and Z is the same and since both lead to equivalent heuristics we have chosen Q for its computational simplicity. For refined work we might adopt Z.

and extends this representation. We distinguish such heuristics from those discussed previously which extend R directly. We call these two ways of looking at R, the extensional and intensional interpretations.

This situation is clarified if we look at the probabilistic matrix:

| | $I_1$ | $I_2$ | | $I_m$ | R |
|---|---|---|---|---|---|
| $D_1$ | $w_{11}$ | $w_{12}$ | . . . | $w_{1m}$ | $w_1(R)$ |
| $D_2$ | $w_{21}$ | $w_{22}$ | . . . | $w_{2m}$ | $w_2(R)$ |
| . | . | . | . . . | . | . |
| . | . | . | . . . | . | . |
| . | . | . | . . . | . | . |
| $D_n$ | $w_{n1}$ | $w_{n2}$ | . . . | $w_{nm}$ | $w_n(R)$ |

We have adjoined to the matrix a column giving the w-values for a given request R. Since the non-zero values in this column characterize the documents that are retrieved this represents the extensional interpretation of R. The values themselves in the R-column can be thought of as a measure of closeness between R and the documents. The matrix $\lceil w_{ij} \rceil$ itself gives a representation of the document space and the index space. To get the intensional interpretation of R into the schema we use the following device: Write R in so-called "distinguished disjunctive normal form"--this is a disjunction of conjunctions in which either $I_j$ or its negation occurs. For example, in the space of three tags $I_1, I_2, I_3$ the request $R = I_1 . I_2$ can be written

$$R = I_1 . I_2 . I_3 \vee I_1 . I_2 . \overline{I}_3 .$$

Having done this, we can represent each conjunction that occurs in R by a vector whose $j^{th}$ component $e_j$ is 1 if $I_j$ is in the conjunction, and 0 otherwise. Thus R is represented by a bundle of such vectors and we have a matrix whose rows are these vectors:

$$B = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1m} \\ e_{21} & e_{22} & \cdots & e_{2m} \\ \cdot & & & \\ e_{s1} & e_{s2} & \cdots & e_{sm} \end{bmatrix}$$

where s is the number of conjunctions that occur in the normal form. We now adjoin this matrix to the probabilistic matrix:

$$\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \cdot & \cdot & \cdots & \cdot \\ w_{n1} & w_{n2} & \cdots & w_{nm} \\ e_{11} & e_{12} & \cdots & e_{1m} \\ \cdot & \cdot & \cdots & \cdot \\ e_{s1} & e_{s2} & \cdots & e_{sm} \end{bmatrix}$$

intensional representation of r

A simple heuristic can be generated from this schema by replacing the non-zero w-values by 1's, thus obtaining the binary library matrix. The square of the Pythagorean distance between rows gives the number of positions in which the rows differ. We say two rows are 0-away, 1-away, etc.,[1] depending on the value of the square distance. We see

[1] The notion of distance between documents is analogous to the notion of distance between codes as discussed in the theory of error correcting codes.

that R will retrieve document $D_1$ if and only if there is a row in B that is 0-away from the $D_1$ row of the binary matrix. (In fact, there will be at most one such row.) In this case we say that $D_1$ is 0-away from R. We can enlarge the class of retrieved documents by considering also documents that are 1-away from R, 2-away, etc.

Enlarging the class by this method is not completely satisfactory. We would really like to introduce these notions into a (generalized) relevance number computation. That is to say, we would like to combine heuristics in such a way that documents with associated ranking numbers are retrieved, not just classes of documents. We would also like to use the values $w_i(R)$ in the computation.

First we note that the Pythagorean distance between two rows of the probabilistic matrix gives a measure of dissimilarity of information content (as well as dissimilarity of distribution of information) between documents corresponding to these rows. Call this distance "$\triangle(D_i, D_j)$". We can use this distance function to compute the distance of any document from the class C of documents retrieved by the basic selection process. This is all the theory required to implement formula (4) (section 2.2).[1]

Next is the problem of computing the (generalized) relevance number. There are infinitely many possibilities here and which is "best" is still an open problem. However, an extremely natural one arises as follows: We have pointed out that the values $w_i(R)$ in the R-column of the probabilistic matrix can be considered as a measure of closeness

[1] Another measure of dissimilarity is to take, not the Pythagorean distance but the sum of the absolute values of the differences between corresponding entries; i.e., $\sum_k |w_{ik} - w_{jk}|$

between R and the documents. To combine these values with $\Delta(D_i, D_j)$ we convert closeness to "distance" by some device such as considering the negative of the logarithm of $w_i(R)$. We define

$$d(R, D_i) = - \log w_i(R). \tag{13}$$

$D_i$ will be retrieved by R if and only if $d(R, D_i)$ is finite; thus this characterizes the class of retrieved documents. Now take that document $D_i$ in the class of retrieved documents such that $\Delta(D_i, D_j)$ is a minimum.[1]

Then we take

$$g(R, D_j) = \sqrt{\Delta^2(D_i, D_j) + \log^2 w_i(R)} \tag{14}$$

as the measure of "distance"[2] between R and $D_j$. Note that if $D_j$ is a retrieved document, then $\Delta(D_i, D_j)$ is zero and

$$g(R, D_j) = - \log w_j(R).$$

Furthermore, if $D_j$ has not been retrieved (initially)

$$g(R, D_j) > - \log w_i(R),$$

where i is the accession number of the document nearest to $D_j$. Thus the ranking by the g-function will always put an adjoined document below its associated document in the class C. We may now finish the computation by subtracting the logarithm of the a priori probability of a document from its g-value. (Analogous to multiplying $w_i(R)$ by $P(A, D_i)$ to obtain the relevance number.) The final heuristic then, is to choose a suitable cut off point in the list of adjoined documents--taking only those with (generalized) relevance number less than some specified value.

---

[1] If $D_i$ is not unique choose the one in the minimal set with the largest $w_i(R)$.

[2] Again (see note 1, p 47) it might be preferable to take

$$g(R, D_j) = \Delta(D_i, D_j) + d(R, D_i)$$

Although simple in theory the above heuristic leads to laborious computations. Considerable simplication results if we restrict all computations to the columns of the probabilistic matrix corresponding to those index terms mentioned in R.

## 2.8 Further Remarks Concerning Search Strategies

We have presented some of the heuristics that appear to have the best possiblity of being useful components of a search strategy. We also have formulated some principles for a general approach to the problem of automatic elaboration of the selection process. Let us now illustrate these ideas by constructing an over-all search strategy.

First we list the variables involved:

1. Input
   (a) The request R
   (b) The request weights
2. The Probabilistic Matrix $[w_{ij}]$
   (a) Similarity measures between documents. (e.g., $\triangle$-values)
   (b) Significance measures for index terms (An index term applied to every document in the library will have no significance, while an index term applied to only one document will be highly significant. Thus significance measures are related to the "extension number" for each term; i.e., to the number of documents tagged with the term--the smaller this number, the greater the significance of the index term.)
   (c) "Closeness" measures between index terms (e.g., Q-values)
3. The A Priori Probability Distribution
4. Output (by means of the basic selection process; i.e., the logical match plus Bayes' schema with all of its ramifications and refinements).

(a) The class of retrieved documents; call this 'C'.

(b) n, the number of documents in C.

(c) Relevance numbers.

5. Control Numbers

(a) $n_o$, the maximum number of documents that we wish to retrieve.

(b) Relevance number control; e.g., we may ignore documents with relevance number less than a specified value.

(c) Generalized relevance number control. (Similar to the above but this applies to the computation described in section 2.7)

(d) Request weight control; i.e., we elaborate on index terms in the request if their request weight is higher than some specified value.

(e) Significance number of index term control; i.e., we give index terms of certain significance (defined in terms of their extensions) special attention.

6. Operations

(a) Basic selection process, denote this by "f".

(b) Elaboration of the request by using "closeness" in the request space. Denote this by "H". Thus the operation H will transform the request R into a new request R'. More precisely H is the heuristic: elaborating the index terms in R with request weights greater than the request weight control number and/or index term significance greater than a specified value.

(c) Adjoining new documents to the class of retrieved documents by using "distance" in the document space. Denote this by "h". Thus the operation h will transform the class C of retrieved documents into a new class, say D. Move precisely, h is the heuristic: trim C to documents having relevance number greater than the control number

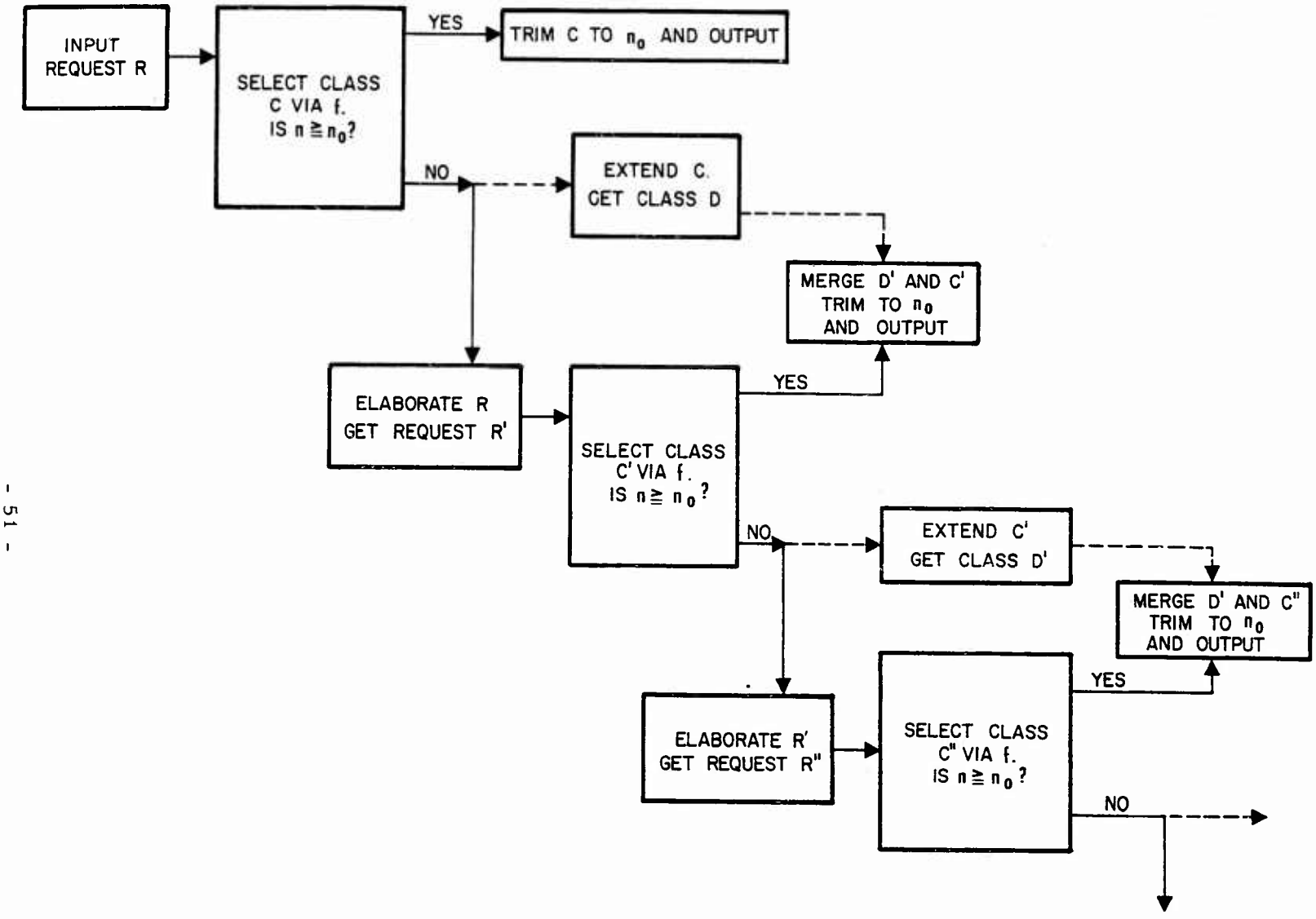



DIAGRAM 3

A SEARCH STRATEGY BASED ON HEURISTICS OF ELABORATION AND EXTENSION

and then annex to C all of the documents with generalized relevance number in a certain range.

(d) Merge: any merging operation between two classes; e.g., forming their intersection, their union, trimming by using relevance number and then forming union, etc.

Next we combine these to obtain the strategy shown in diagram 3. This strategy is to be regarded as a particularly simple example, its goal to obtain a specified number of documents (say $n_o$) having the best chance of satisfying the request. Thus the decision to elaborate centers on answering the question: Is the number of documents selected greater than or equal to $n_o$? In the diagram we refer to the heuristic H as simply "elaborate the request". The actual transfer function H involves using control numbers to limit the elaboration. Furthermore these control numbers can be varied from one application of H to the next. Similarly we refer to the heuristic h as simply "extend the class"; but we point out that this too involves control numbers. Finally a word about the classes C, C', D, etc. These are actually lists of documents ranked by relevance numbers. Thus the instruction "trim C to $n_o$" means "cut/off the list to the $n_o$ documents with highest relevance number". The output of the system will be an ordered list of document accession numbers.

# PART III.

## THE EXPERMENTAL RESULTS

## (SUMMARY)

In Part III we describe some experiments that were designed and executed in order to provide data for evaluating the effectiveness of the techniques of Probabilistic Indexing. The discussion of Part II indicates that there are two basic hypotheses that we wish to verify. The first hypothesis asserts that the relevance number that we compute for each document, given a request is, in fact, a measure of the probable relevance of the document. The second hypothesis asserts that the automatic elaboration of the selection process does, in fact, produce relevant documents which are not selected by the original request.

Section 1 discusses the experimental set-up; i.e., the library, the indexing system, the weights, the testing procedure, etc. Section 2 provides the data and discussion in support of the hypothesis concerning the relevance function and we find that the results do support the hypothesis. Section 3 provides the data and discussion in support of the hypothesis concerning the selection process and we find that the results do support the hypothesis.

# 1. THE EXPERIMENTAL SET-UP

## 1.1 Initial Remarks

The jumping off point for our approach to automatic information retrieval was the recognition that the core of the problem is that of adequately identifying the information content of documentary data. There is an uncertainty in the relationship between the tags that are used to index documents and the subjects that they denote and this is the cause of inadequate retrieval of desired information. Using the analogy of going from an incoming document to its set of index tags as going from a selected message to a received message over a noisy channel, we recognized the effect of semantic noise and accordingly a technique for handling it statistically. Given this analogy the problem was to select the proper schema from the calculus of probability to allow for the inverse inference from requested index terms to most probably relevant document. This line of reasoning thus led us to the notion of weighting index tags and using Bayes' Theorem to provide a function to measure the degree of relevance between an arbitrary request and any of the documents selected by the request. Further analysis led us to the notion of automatically elaborating upon the request in the most probable direction so as to improve the selection of relevant documents.

Given the fundamental notions of Probabilistic Indexing and a logico-mathematical explication (presented in Part II) with which to back up our intuitive understanding of the problem, let us now raise the question of justification. That is to say, to what extent does our probabilistic analyses of the library problem guarantee that retrieval effectiveness will be improved. Clearly, the only real justification for Probabilistic Indexing is success; i.e., if the technique improves retrieval effectiveness then the system is justified and if not, not. Therefore, given the basic methods of Probabilistic Indexing a natural next step is to conduct some actual library experiments in order to measure its degree of success in improving retrieval effectiveness. In the following sections we shall describe the design, execution and results of some actual experiments.

An evaluation of these empirical results provides good evidence in favor of the "theory" of Probabilistic Indexing.

Let us point out at this time, that the value of actual library experiments goes beyond a mere campaign to evaluate the techniques of Probabilistic Indexing. Actual library experimentation provides an excellent tool by means of which we can refine and extend the methods and techniques that constitute Probabilistic Indexing. We have formulated already what we feel are excellent approximate solutions to some of the major problems and these notions now must be verified and refined, where necessary, on the basis of actual experience. Just as the physicist requires such tools as, for example, a linear accelerator in order to empirically verify and suggest new notions relative to nuclear physics, so also the "library scientist" requires the counterpart of the linear accelerator; viz., a library with which he can work and control. Just as an experiment in physics represents a set of questions that a physicist asks of Nature, so also the library scientist needs an experimental library to which he can ask questions about the nature of information identification indexing, searching, etc.; and thereby to obtain answers on the basis of which to refine his original questions and provide insights into the evidence relative to library problems and their solutions. Hopefully, our explications can be refined so as to provide us with a good first approximation to a fundamental theory of literature identification, indexing, searching and retrieval.

## 1.2 The Experimental Library

A collection of articles from Science News Letter[1] formed the library for our experiments. The Science News Letter is a weekly summary of current events in science and the articles cover a wide range of subjects ranging from Archaeology to Astronomy, Physics to Physiology, and Medicine to Meteorology. In previous tests made in October of last year we

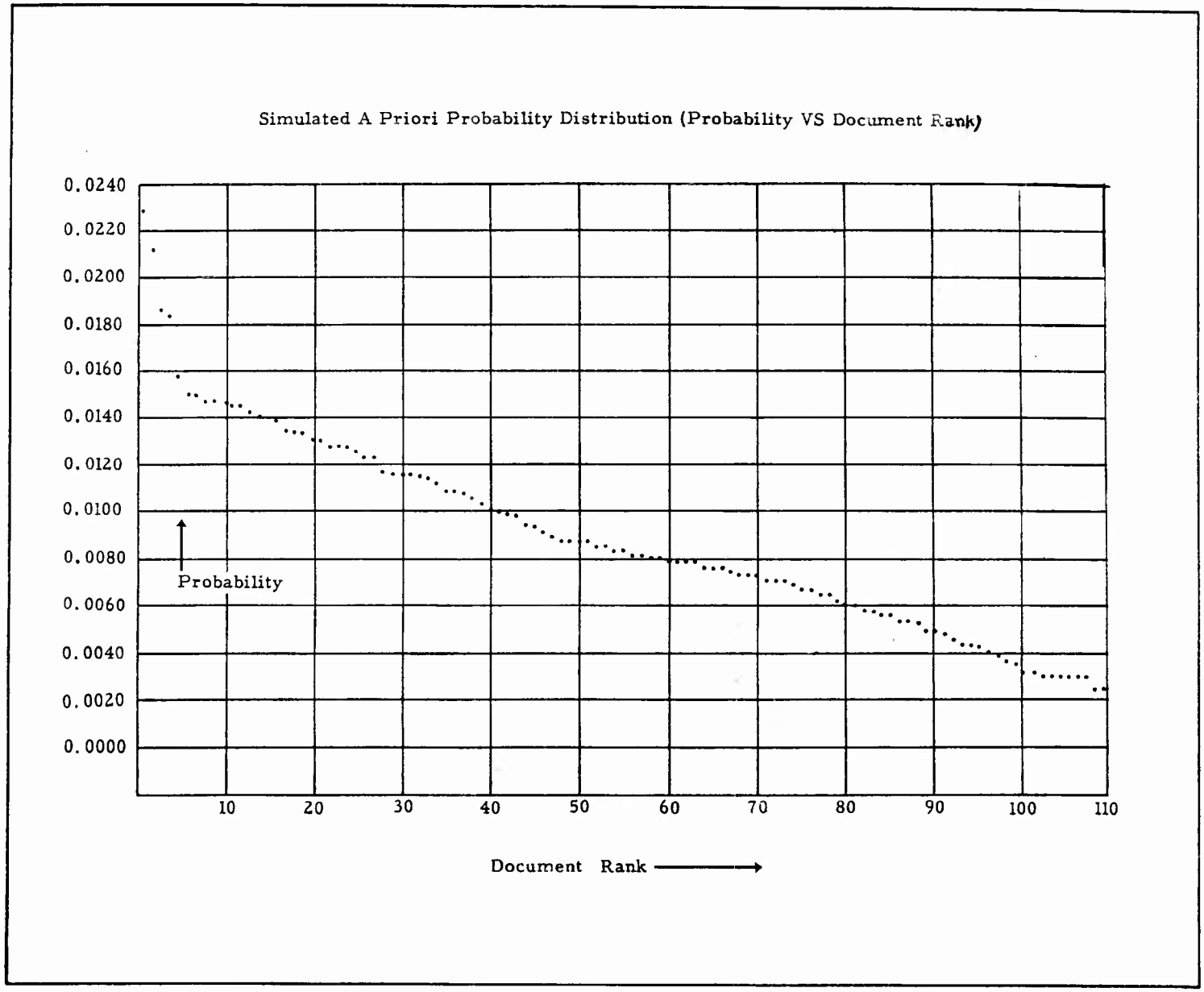
[1] Published by Science Service, Inc., 1719 N Street, N.W., Washington, D.C.

selected 200 articles at random and this collection constituted the experimental library. For the present tests we discarded 90 of the original 200 documents in order to restrict the library to articles dealing with the Physical Sciences. Although 110 is not a large number we feel that our experimental library presented us with most of the relevant and basic problems that would be found in a "real" library and yet still be of a manageable size.

Our choice of articles from Science News Letter for inclusion in the experimental library was dictated to a large extent by the fact that these articles are relatively brief, pithy, clearly written, interesting, timely, uncomplex and easy to index by non-experts. This made not only the indexing but the subsequent evaluation of retrieval documents a reasonably uncomplicated task. Since this experimental library was one without a previous history of usage there were no statistics on the a priori probability of document usage and consequently these statistics had to be simulated as discussed in Part II, 1.12. Graph 1 shows the simulated non-linear distribution.

## 1.3 The Indexing System

The indexing system, again, refers to the class of tags that are used to identify both the content of the documents and the requests and thus it is the language common to both "sides" of the library. Since the methods of Probabilistic Indexing are applicable to any indexing system we were not limited in our choice of a set of tags to be used for the experimental library. The only constraint was that the number of tags in the index list be comparable with the size of the library. Instead of "truncating" an existing index system and using its tags to index the documents of the experimental library, we adopted the following procedure: Each document of the library was read and the key content bearing words were selected and listed. There were a total of 577 different keywords in the list. These words were sorted into categories on the basis of their meanings. It turned out that the keywords (as they are called) could be sorted into 47 fairly well-defined categories. In many cases a particular keyword would


Simulated A Priori Probability Distribution (Probability VS Document Rank)
0.0240
0.0220
0.0200
0.0180
0.0160
0.0140
0.0120
0.0100
0.0080
0.0060
0.0040
0.0020
0.0000
Probability
10
20
30
40
50
60
70
80
90
100
110
Document Rank

GRAPH 1

belong to more than one category, consequently there were 919 occurrences of the keywords in the 47 categories. The names of the 47 categories are listed in Table 2 and these names became the tags that constituted the index term list for the experimental library.

These 47 index terms were then assigned to the documents by working backwards as follows: For each category we determined which keywords it contained and each document which contained the keyword in question, was coordinated in the corresponding category. That is to say, given the categories, the keywords in each category and the documents associated with each of the keywords, we then were able to determine which documents should be coordinated with each category and thus the documents were indexed by assigning to each the names of the corresponding categories. This is clarified in Diagram 4 which shows the relationships between documents and keywords, keywords and categories, and, therefore, documents and categories. Table 3 shows the number of keywords that were associated with each of the 47 categories and also the number of documents associated with each of the 47 categories. Graph 2 shows the distribution of the frequency with which the index terms were used, plotted against their rank.

### 1.4 The Assignment of Weights

Having assigned the index terms to the documents, Probabilistic Indexing requires that we indicate the degree with which each tag holds for the document by assigning weights to the index terms. In order to assign the corresponding weight each document was reread and then the indexer decided for each of the tags coordinated to each document, the degree with which it held. We had decided previously that a reasonable range of values for the weights was eight, ranging from 1/8 to 8/8. In order to aid the indexer in obtaining a consistent assignment of weights, rough weighting rules were formulated and these are shown in Table 4. Table 5 shows the distribution of weights for each of the 47 index terms. A portion of the Probabilistic Library matrix is shown in the Table p. 65.

1. Aerodynamics and Aviation
2. Agriculture
3. Animals (including birds, fish, and reptiles)
4. Archaeology
5. Astronomy
6. Atmosphere
7. Atomic Physics
8. Biology
9. Chemistry
10. Communications
11. Computers
12. Defense and Warfare
13. Electronics
14. Engineering
15. Engines
16. Food
17. Geography
18. Geology
19. Geophysics
20. Health and Safety
21. History
22. Machinery
23. Man
24. Materials
25. Mathematics
26. Measurement
27. Missiles and Rockets
28. Mystery, Myths and Problems
29. Nature
30. Navigation
31. Paleontology
32. Physical Quantities
33. Physics
34. Plants
35. Political or government groups or functions
36. Power
37. Predictions
38. Psychology
39. Research
40. Satellites
41. Social Sciences
42. Space Travel
43. Teaching - Education
44. Time
45. Tools
46. Transportation
47. Weather

TABLE 2

INDEX TERMS
DERIVED FROM KEYWORDS

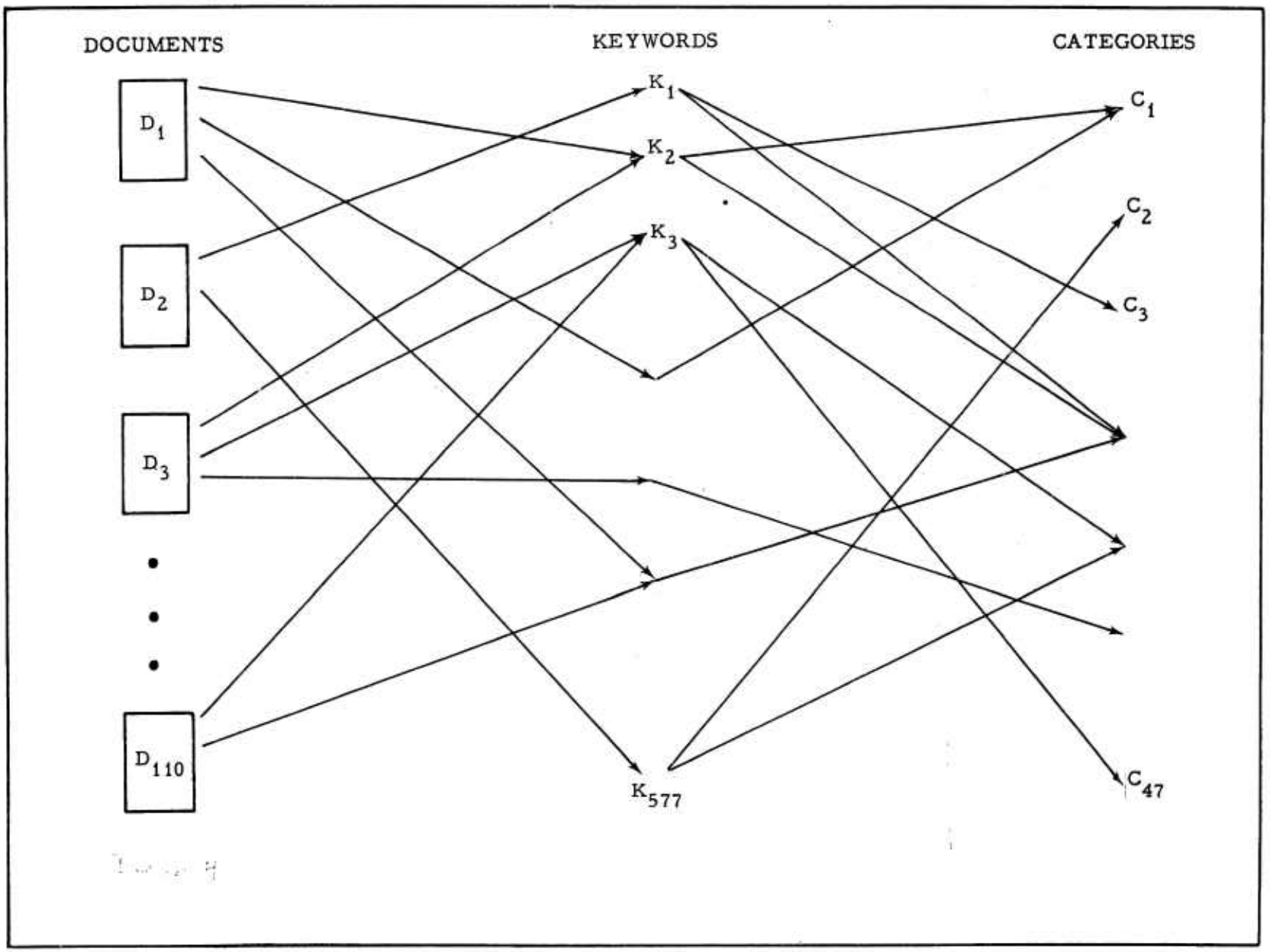
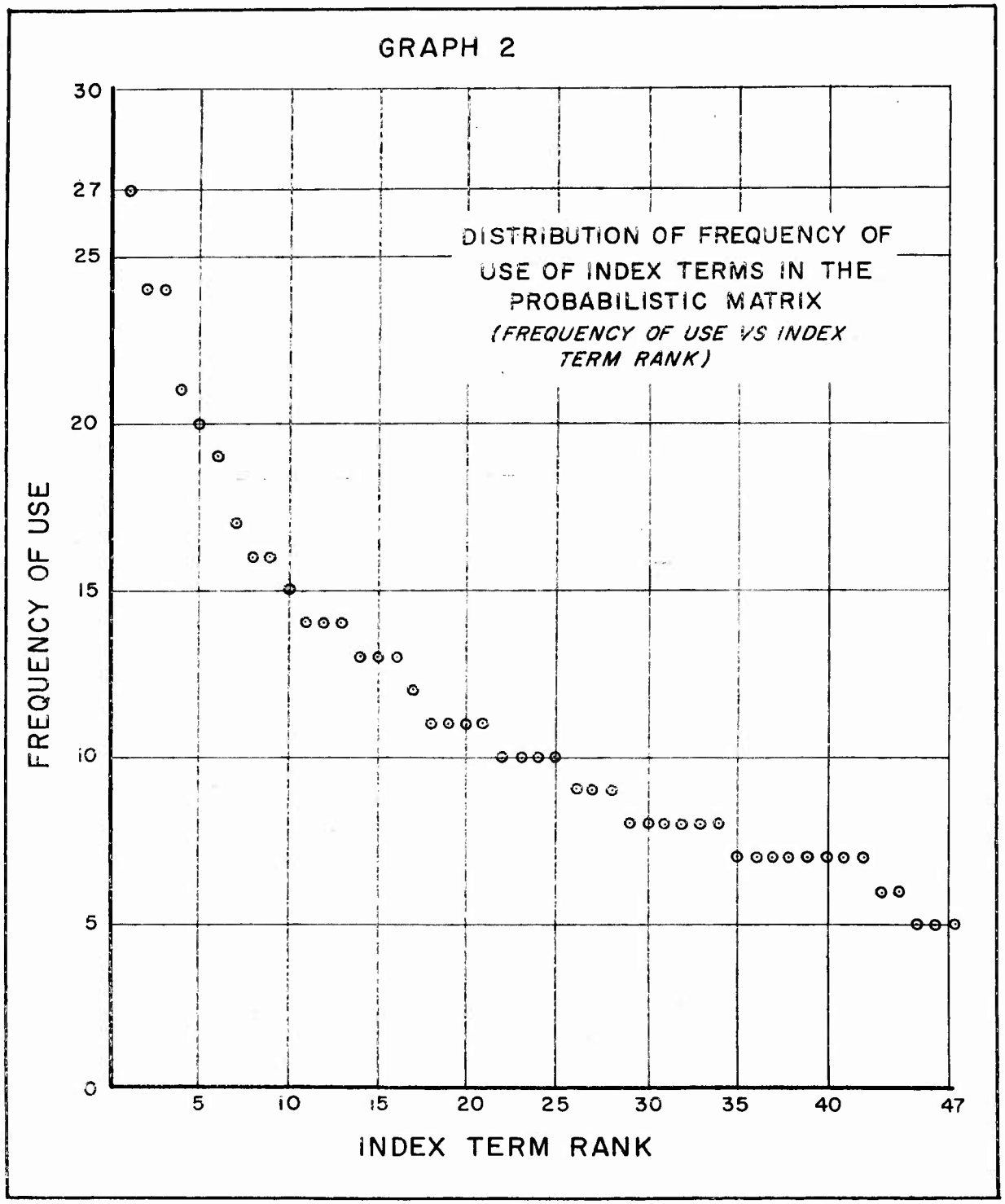



DIAGRAM 4. RELATIONSHIPS BETWEEN DOCUMENTS, KEYWORDS AND CATEGORIES

| | | Number of documents | Number of keywords | | | Number of documents | Number of keywords |
|---|---|---|---|---|---|---|---|
| 1 | Aerodynamics and Aviation | 17 | 30 | 25 | Mathematics | 7 | 14 |
| 2 | Agriculture | 8 | 12 | 26 | Measurement | 8 | 35 |
| 3 | Animals (including birds, fish, and reptiles) | 13 | 27 | 27 | Missiles and Rockets | 9 | 21 |
| 4 | Archaeology | 8 | 12 | 28 | Mystery, Myths and Problems | 7 | 7 |
| 5 | Astronomy | 13 | 29 | 29 | Nature | 9 | 9 |
| 6 | Atmosphere | 13 | 22 | 30 | Navigation | 7 | 10 |
| 7 | Atomic Physics | 14 | 35 | 31 | Paleontology | 5 | 9 |
| 8 | Biology | 11 | 19 | 32 | Physical Quantities | 16 | 28 |
| 9 | Chemistry | 15 | 23 | 33 | Physics | 20 | 30 |
| 10 | Communications | 10 | 25 | 34 | Plants | 11 | 28 |
| 11 | Computers | 7 | 12 | 35 | Political or government groups or functions | 27 | 43 |
| 12 | Defense and Warfare | 11 | 21 | 36 | Power | 13 | 21 |
| 13 | Electronics | 14 | 16 | 37 | Predictions | 5 | 6 |
| 14 | Engineering | 11 | 14 | 38 | Psychology | 6 | 9 |
| 15 | Engines | 8 | 15 | 39 | Research | 7 | 8 |
| 16 | Food | 7 | 11 | 40 | Satellites | 10 | 17 |
| 17 | Geography | 21 | 36 | 41 | Social Sciences | 8 | 17 |
| 18 | Geology | 8 | 12 | 42 | Space Travel | 7 | 7 |
| 19 | Geophysics | 10 | 11 | 43 | Teaching - Education | 5 | 17 |
| 20 | Health and Safety | 13 | 15 | 44 | Time | 7 | 12 |
| 21 | History | 10 | 21 | 45 | Tools | 6 | 9 |
| 22 | Machinery | 24 | 29 | 46 | Transportation | 16 | 27 |
| 23 | Man | 20 | 29 | 47 | Weather | 9 | 21 |
| 24 | Materials | 24 | 38 | | | | |

USE OF INDEX TERMS


GRAPH 2
DISTRIBUTION OF FREQUENCY OF USE OF INDEX TERMS IN THE PROBABILISTIC MATRIX
(FREQUENCY OF USE VS INDEX TERM RANK)
FREQUENCY OF USE
INDEX TERM RANK
0
5
10
15
20
25
27
30
5
10
15
20
25
30
35
40
47

| WEIGHT | DESCRIPTION | WHEN USED |
|---|---|---|
| 8/8 | Major Subject | The term is highly specific and covers an entire major subject of the document. |
| 7/8 | Major Subject | The term is specific and covers most of a major subject of the document. |
| 6/8 | More Generic Subject | The term is too broad and covers a major subject. |
| 5/8 | Other Important Terms | Terms that would be used in a binary indexing but not a major subject. |
| 4/8 | Less Generic Subject | The term relates to but is too narrow to cover a major subject. |
| 3/8 | Minor Subject | Includes such terms as relate to results of experiments, intermediate methods, possible uses, etc. |
| 2/8 | Other Subjects | Other relevant tags. |
| 1/8 | Barely relevant | Subjects classifier would not want to use but feels that some users might consider them relevant. |

TABLE 4

A GUIDE FOR THE ASSIGNMENT OF WEIGHTS

| | | WEIGHTS | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 8/8 | 7/8 | 6/8 | 5/8 | 4/8 | 3/8 | 2/8 | 1/8 | Totals |
| 1. | Aerodynamics and Aviation | 1 | 5 | 3 | 1 | 2 | 3 | 2 | 0 | 17 |
| 2. | Agriculture | 1 | 2 | 2 | 0 | 2 | 1 | 0 | 0 | 8 |
| 3. | Animals | 1 | 3 | 6 | 1 | 3 | 1 | 1 | 1 | 13 |
| 4. | Archaeology | 1 | 2 | 4 | 0 | 0 | 0 | 0 | 1 | 8 |
| 5. | Astronomy | 3 | 2 | 4 | 1 | 1 | 1 | 0 | 0 | 12 |
| 6. | Atmosphere | 0 | 1 | 2 | 4 | 1 | 2 | 4 | 0 | 14 |
| 7. | Atomic Physics | 3 | 3 | 3 | 0 | 2 | 2 | 1 | 0 | 14 |
| 8. | Biology | 0 | 3 | 5 | 0 | 3 | 0 | 0 | 0 | 11 |
| 9. | Chemistry | 3 | 1 | 4 | 1 | 2 | 2 | 2 | 0 | 15 |
| 10. | Communications | 3 | 2 | 2 | 1 | 0 | 1 | 0 | 1 | 10 |
| 11. | Computers | 2 | 1 | 2 | 1 | 0 | 1 | 0 | 0 | 7 |
| 12. | Defense and Warfare | 3 | 1 | 1 | 0 | 1 | 2 | 0 | 3 | 11 |
| 13. | Electronics | 0 | 1 | 2 | 1 | 3 | 4 | 3 | 0 | 14 |
| 14. | Engineering | 0 | 2 | 4 | 1 | 1 | 1 | 0 | 2 | 11 |
| 15. | Engines | 2 | 3 | 0 | 0 | 2 | 0 | 1 | 0 | 8 |
| 16. | Food | 0 | 1 | 3 | 1 | 1 | 1 | 0 | 0 | 7 |
| 17. | Geography | 0 | 1 | 0 | 4 | 2 | 5 | 4 | 5 | 21 |
| 18. | Geology | 1 | 2 | 1 | 0 | 0 | 2 | 0 | 0 | 8 |
| 19. | Geophysics | 0 | 2 | 1 | 0 | 2 | 0 | 1 | 4 | 10 |
| 20. | Health and Safety | 0 | 2 | 3 | 1 | 4 | 1 | 0 | 2 | 13 |
| 21. | History | 3 | 0 | 4 | 1 | 1 | 1 | 0 | 0 | 10 |
| 22. | Machinery | 1 | 0 | 6 | 1 | 3 | 6 | 4 | 3 | 24 |
| 23. | Man | 0 | 2 | 4 | 3 | 0 | 3 | 5 | 3 | 20 |
| 24. | Materials | 3 | 5 | 3 | 2 | 3 | 5 | 2 | 1 | 24 |
| 25. | Mathematics | 0 | 0 | 3 | 2 | 0 | 2 | 0 | 0 | 7 |
| 26. | Measurement | 0 | 1 | 1 | 0 | 1 | 2 | 2 | 1 | 8 |
| 27. | Missiles and Rockets | 2 | 1 | 1 | 2 | 0 | 2 | 1 | 0 | 9 |
| 28. | Mystery, Myths and Problems | 0 | 0 | 1 | 2 | 1 | 2 | 1 | 0 | 7 |
| 29. | Nature | 0 | 1 | 4 | 1 | 2 | 0 | 1 | 0 | 9 |
| 30. | Navigation | 2 | 1 | 0 | 0 | 0 | 2 | 1 | 1 | 7 |
| 31. | Paleontology | 3 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 5 |
| 32. | Physical Quantities | 0 | 1 | 2 | 1 | 0 | 4 | 7 | 1 | 16 |
| 33. | Physics | 0 | 0 | 4 | 0 | 1 | 6 | 5 | 3 | 19 |
| 34. | Plants | 6 | 4 | 0 | 0 | 0 | 1 | 0 | 0 | 11 |
| 35. | Political or Government etc. | 1 | 2 | 2 | 3 | 2 | 4 | 4 | 9 | 27 |
| 36. | Power | 2 | 2 | 1 | 4 | 3 | 0 | 1 | 0 | 13 |
| 37. | Predictions | 1 | 0 | 1 | 2 | 0 | 0 | 1 | 0 | 5 |
| 38. | Psychology | 3 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 6 |
| 39. | Research | 1 | 0 | 2 | 2 | 0 | 0 | 2 | 0 | 7 |
| 40. | Satellites | 3 | 4 | 3 | 0 | 0 | 0 | 0 | 0 | 10 |
| 41. | Social Sciences | 2 | 2 | 3 | 1 | 0 | 0 | 0 | 0 | 8 |
| 42. | Space Travel | 2 | 0 | 1 | 0 | 0 | 2 | 2 | 0 | 7 |
| 43. | Teaching - Education | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 5 |
| 44. | Time | 0 | 1 | 0 | 0 | 0 | 1 | 4 | 1 | 7 |
| 45. | Tools | 0 | 1 | 0 | 1 | 1 | 1 | 2 | 0 | 6 |
| 46. | Transportation | 1 | 2 | 2 | 2 | 3 | 2 | 2 | 2 | 16 |
| 47. | Weather | 6 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 9 |
| | Totals | 65 | 73 | 103 | 51 | 55 | 76 | 66 | 46 | 535 |

TABLE 5
DISTRIBUTION OF WEIGHTS FOR EACH INDEX TERM

DOCUMENTS

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | ... | 101 | 102 | 103 | 104 | 105 | 106 | 107 | 108 | 109 | 110 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Aerodynamics and Aviation | | | | | | | 2/8 | 6/8 | | | | 7/8 | 3/8 | | 7/8 | | | | | | |
| 2 | Agriculture | | | | | 6/8 | | | | | | | | 6/8 | | | | | | | | |
| 3 | Animals | | | | | | | | | | | | | | | | | | | | | |
| 4 | Archaeology | | | | | | 7/8 | | | | | | | | | | | | | | | |
| 5 | Astronomy | | | | | | | | | | | | | | | | | 8/8 | | | | |
| 6 | Atmosphere | | | | | | | | 2/8 | 3/8 | 5/8 | | 4/8 | | | | | 2/8 | | | | 5/8 |
| 7 | Atomic Physics | 8/8 | 8/8 | 6/8 | | | | 7/8 | 3/8 | | | | | | | | | | | | | |
| 8 | Biology | | | | | | | | | | | | | | | | | | | | 6/8 | |
| 9 | Chemistry | | | | | | | | | | | | | | | | | | | | | |
| 10 | Communications | | | | 3/8 | | | | | | | | | | | | | | | | | |
| 11 | Computers | | | | 5/8 | | | | | | | | 3/8 | | | | | | | | | |
| 12 | Defense and Warfare | | 1/8 | | 1/8 | | | 3/8 | | | | | | | | | | | | | | |
| 13 | Electronics | | 2/8 | | 5/8 | | | | | | | | 2/8 | | | | | | | | | |
| 14 | Engineering | 4/8 | | | | | | | | | | | 1/8 | | | | | | | | | |
| 15 | Engines | | | | | | | 7/8 | 4/8 | | | | | | | | | | | 6/8 | | |
| 16 | Food | | | | | 4/8 | | | | 6/8 | | | | | | | | | | | | |
| 17 | Geography | | | | | | 3/8 | | | | | | | | | | 2/8 | | | | 3/8 | |
| 18 | Geology | | | | | | | | | | | | | 1/8 | | | | | | | | |
| 19 | Geophysics | | | | | | | | | | | | 2/8 | | | | | | | | | |
| 20 | Health and Safety | | | 7/8 | | | | | | | | | | | | 6/8 | | | | | | |
| 21 | History | | | | | 6/8 | | | | | | | | | | | | | | | | |
| 22 | Machinery | 4/8 | | 3/8 | | | | | 1/8 | | | | 1/8 | | | 6/8 | | | 6/8 | 6/8 | | |
| 23 | Man | 1/8 | 1/8 | | 1/8 | | 2/8 | | | 2/8 | | | | | | | | | | | | |
| 24 | Materials | 4/8 | 3/8 | | | | | | | 3/8 | | | | 7/8 | | | | | | 8/8 | | |
| 25 | Mathematics | | 5/8 | | | | | | | | | | | | 6/8 | | | | | | | |
| 26 | Measurement | | | | | | | | | | | | | | | | | | | | | |
| 27 | Missiles and Rockets | | | | | | | 3/8 | 6/8 | | 7/8 | | | | | | | | | | | |
| 28 | Mystery, Myths and Problems | | | | | | 5/8 | | | | | | | | | | | | | | | |
| 29 | Nature | | | | | | | | | | 6/8 | | | | | | | | | | | |
| 30 | Navigation | | | | | | | | 2/8 | | | | | | | 8/8 | | | | | | |
| 31 | Paleontology | | | | | | 4/8 | | | | | | | | | | | | | | | |
| 32 | Physical Quantities | | | | | | | | 3/8 | | | | 2/8 | 1/8 | | | | | | | | |
| 33 | Physics | | 6/8 | | | | | | | | 3/8 | | | 2/8 | | | | | | | | |
| 34 | Plants | | | | | 7/8 | | | | | | | | 8/8 | | | | | | | 8/8 | |
| 35 | Political or Government etc. | 1/8 | 1/8 | | 1/8 | | 1/8 | 6/8 | | | | | | | | | | | | | 1/8 | |
| 36 | Power | | 2/8 | | | | | 4/8 | 5/8 | | | | | | | | | | | | | |
| 37 | Predictions | | | | | | | | | | | | | | | | | | | | | |
| 38 | Psychology | | | | 6/8 | | | | | | | | | | | | | | | | | |
| 39 | Research | | | | | 2/8 | | | | | | | 2/8 | | | | | | | | | |
| 40 | Satellites | | | | | | | | | | | | | | | | | | | | | 7/8 |
| 41 | Social Sciences | | | | | | | | | | | | | | | | 8/8 | | 7/8 | | | |
| 42 | Space Travel | | | | | | | | 3/8 | 6/8 | | | | | | | | | | | | |
| 43 | Teaching - Education | | | | | | | | | | | | | | 7/8 | | | | | | | |
| 44 | Time | | | | | | | | | | | | | | 1/8 | | | 2/8 | | | | |
| 45 | Tools | | | 3/8 | | | | | | 7/8 | | | 2/8 | | | | | | | | | |
| 46 | Transportation | | | | | | | 3/8 | 4/8 | | | | 1/8 | | | 5/8 | | | 6/8 | 4/8 | | 4/8 |
| 47 | Weather | | | | | | | | | | | | 8/8 | 7/8 | | | | | | | | |

A PORTION OF THE PROBABILISTIC LIBRARY MATRIX

Given the 47 index terms we computed all the conditional probabilities; i.e., the probabilities that given a document tagged with $I_j$, it will be tagged also with $I_k$. This conditional probability matrix is shown in the Table p. 67. The Table p. 68 lists for each of the index terms the term for which it has the highest forward conditional probability; i.e., for each $I_j$ it shows that $I_k$ which makes $P(I_j, I_k)$ a maximum. In order to show more graphically the connections between these terms we have included the "map" in Diagram 5. The Table p. 70 shows the most highly correlated inverse conditional probabilities; i.e., for each $I_j$ it shows that $I_k$ which makes $P(I_k, I_j)$ a maximum. The Table p. 71 is the matrix which shows the coefficients of association that each index term has for every other term and in the Table p. 72 the most highly correlated terms (in the sense of coefficient of association) are shown for each of the 47 index terms.

## 1.5 The Testing Procedure

Our procedure for empirically testing the notions of Probabilistic Indexing was rather straightforward and unsophisticated and can be described as follows: At random we selected 34 documents from the experimental library and from these "selected" documents, we formulated 40 different questions. These questions were not so narrow that they demanded some specific piece of data for the proper answer, but, rather, they were more general and of the type; "Find information on the use of atmospheric energy for satellite propulsion", rather than, "Find the specific gravity of beryllium".

We then chose five intelligent cooperative technical members of the Company and asked them to act as test subjects. Each was given a list of eight questions and a list of the library index terms. They were briefed as to the nature of the library and they were asked to formulate a library request for information on each of the eight questions. We requested that they attempt to have the "correct" document retrieved and as little else

| Index Term | Implies this Term with the weight shown | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aerodynamics and Aviation | 1 | 1 | .059 | .059 | - | - | .236 | .118 | - | .118 | .118 | .059 | .176 | .118 | .353 | .353 | - | .118 | .059 | .118 | - | | .589 | .059 | .176 | | .059 | .176 | .118 | .059 | .236 | - | .294 | .236 | .059 | .236 | .294 | .118 | .059 | .176 | .118 | .059 | .059 | - | .059 | .059 | .646 | .353 |
| Agriculture | 2 | .125 | 1 | .125 | - | - | - | - | .125 | .250 | - | - | - | - | .125 | - | .375 | .125 | .125 | .125 | - | .125 | - | .125 | .625 | - | - | - | - | .375 | - | - | .125 | .125 | .625 | .375 | - | - | .125 | - | - | - | - | .250 | .125 | - | .125 |
| Animals | 3 | .077 | .077 | 1 | .231 | - | - | - | .385 | .077 | - | - | - | - | - | - | .154 | .154 | - | .154 | .154 | .231 | .077 | .154 | .231 | - | .077 | - | .077 | .231 | - | .308 | .077 | .077 | .154 | .077 | - | - | .077 | - | - | - | .077 | .077 | .077 | - | - |
| Archaeology | 4 | - | - | .375 | 1 | - | - | .250 | .125 | - | .250 | - | - | - | - | - | - | .750 | - | .250 | - | .875 | .125 | .375 | .250 | - | - | - | .125 | - | .500 | - | .125 | .125 | .625 | .125 | - | - | - | - | - | - | .125 | .125 | - | - |
| Astronomy | 5 | - | - | - | - | 1 | .231 | .154 | .154 | .077 | .077 | - | - | .231 | - | - | - | .154 | .231 | - | - | - | .077 | - | .077 | - | .154 | - | - | .077 | .154 | - | .154 | .308 | - | .231 | - | .077 | - | - | .308 | - | .231 | - | .077 | - | - | - |
| Atmosphere | 6 | .308 | - | - | - | .231 | 1 | .077 | - | .154 | .077 | .077 | - | .231 | .154 | .154 | .077 | - | - | .231 | .077 | - | .308 | .154 | .154 | .077 | - | .231 | - | .077 | .077 | - | .308 | .231 | - | .154 | .077 | .077 | - | .154 | .231 | - | .231 | - | .077 | .154 | .308 | .154 |
| Atomic Physics | 7 | .143 | - | - | .143 | .143 | .071 | 1 | - | .071 | - | - | .357 | .071 | .143 | .143 | - | .286 | .143 | .071 | .214 | .071 | .357 | .143 | .428 | .071 | - | .286 | - | - | .143 | - | .143 | .428 | .071 | .643 | .428 | - | - | - | .071 | - | .214 | .071 | .071 | .071 | .143 | - |
| Biology | 8 | - | .091 | .455 | .091 | .182 | - | - | 1 | .364 | - | - | - | - | - | - | .182 | .182 | - | .091 | .182 | .091 | - | .091 | .182 | - | - | - | - | .182 | - | .182 | - | - | .364 | .091 | - | - | - | - | - | - | - | .091 | - | - | - | - |
| Chemistry | 9 | .133 | .133 | .067 | - | .067 | .133 | .067 | .267 | 1 | - | - | .133 | .067 | .133 | .067 | - | .067 | .067 | .133 | - | .067 | .267 | .133 | .667 | - | .067 | - | .133 | .133 | - | - | .067 | .200 | .133 | .133 | .267 | - | - | - | - | - | - | - | - | .167 | - |
| Communications | 10 | .200 | - | - | .200 | .100 | .100 | - | - | - | 1 | .100 | .200 | .400 | - | - | - | .200 | - | .100 | .100 | .200 | .400 | .500 | .100 | - | - | - | - | .100 | - | .200 | .300 | - | .400 | - | .100 | - | - | .200 | - | - | - | - | - | .100 | .100 |
| Computers | 11 | .142 | - | - | - | - | .142 | - | - | - | .142 | 1 | .142 | .714 | .142 | - | - | - | - | .285 | - | - | .428 | .142 | - | .428 | .142 | - | .142 | - | - | - | .285 | - | - | .142 | - | .142 | .142 | .285 | - | - | - | - | .285 | .152 | .285 |
| Defense and Warfare | 12 | .273 | - | - | - | - | - | .455 | - | .182 | .182 | .091 | 1 | .182 | .091 | .182 | - | .182 | .091 | .091 | .182 | - | .273 | .364 | .273 | .091 | - | .182 | .091 | - | .091 | - | - | .182 | - | .546 | .364 | - | .091 | - | - | .091 | - | - | - | .364 | - |
| Electronics | 13 | .143 | - | - | - | .214 | .214 | .071 | - | .071 | .286 | .357 | .143 | 1 | .143 | - | - | - | .071 | .286 | - | - | .428 | .214 | .143 | .143 | .143 | - | - | - | .071 | - | .286 | .428 | - | .286 | .143 | .214 | .071 | .071 | .286 | - | .071 | - | - | .071 | .071 | .214 |
| Engineering | 14 | .546 | .091 | - | - | - | .182 | .182 | - | .182 | - | .091 | .091 | .182 | 1 | .273 | .091 | .182 | - | .182 | .091 | - | .637 | .182 | .455 | - | - | - | .091 | .091 | - | - | .273 | .273 | - | .091 | .182 | - | - | .182 | - | - | - | .182 | .091 | .182 | .182 |
| Engines | 15 | .750 | - | - | - | - | .250 | .250 | - | .125 | - | - | .250 | - | .375 | 1 | - | .250 | - | - | .125 | - | .500 | .250 | .250 | - | - | .375 | .250 | - | .125 | - | .250 | .125 | - | .250 | .750 | - | - | - | .250 | - | .125 | - | .125 | - | .625 | .125 |
| Food | 16 | - | .428 | .285 | - | - | .142 | - | .285 | - | - | - | - | - | .142 | - | 1 | .142 | - | .142 | .285 | - | - | .285 | .428 | - | - | - | - | .142 | - | - | - | - | .428 | - | - | - | - | .142 | - | - | .142 | - | .142 | .142 | - | - |
| Geography | 17 | .895 | .047 | .095 | .286 | .095 | - | .190 | .095 | .047 | .095 | - | .095 | - | .095 | .095 | .047 | 1 | .143 | .190 | .047 | .333 | .095 | .143 | .190 | - | .190 | .095 | .095 | .190 | .095 | .095 | .190 | .190 | .095 | .714 | .143 | - | - | .095 | .143 | .095 | .047 | - | .190 | - | .047 | .095 |
| Geology | 18 | .125 | .125 | - | - | .375 | - | .250 | - | .125 | - | - | .125 | .125 | - | - | - | .375 | 1 | .375 | - | - | .125 | - | .375 | - | .125 | - | .125 | .125 | .125 | - | .250 | .625 | .125 | .250 | .250 | - | - | .125 | - | - | .125 | - | - | - | - | .125 |
| Geophysics | 19 | .200 | .100 | .200 | .200 | - | .300 | .100 | .100 | .200 | .100 | .200 | .100 | .400 | .200 | - | .100 | .400 | .300 | 1 | - | .100 | .300 | .300 | .300 | .100 | .100 | - | - | .100 | - | .100 | .300 | .300 | .100 | .200 | .200 | - | - | .300 | .200 | - | - | - | .100 | .100 | .200 |
| Health and Safety | 20 | .384 | - | .154 | - | - | .077 | .231 | .154 | - | .077 | - | .154 | - | .077 | .077 | .154 | .077 | - | - | 1 | - | .308 | .154 | .077 | .077 | - | .154 | .077 | .077 | .154 | - | .154 | .231 | .077 | .077 | .154 | .077 | .154 | - | - | - | .077 | - | - | - | .384 | .077 |
| History | 21 | - | .100 | .300 | .700 | - | - | .100 | .100 | .100 | .200 | - | - | - | - | - | - | .700 | - | .100 | - | 1 | - | .300 | .200 | - | .100 | - | .100 | .200 | - | .400 | .100 | - | .100 | .500 | - | - | - | - | - | .100 | .100 | - | .100 | - | - |
| Machinery | 22 | .416 | - | .041 | .041 | .041 | .166 | .208 | - | .166 | .166 | .125 | .125 | .250 | .291 | .166 | - | .083 | .041 | .125 | .166 | - | 1 | .125 | .375 | .083 | .041 | .083 | .083 | .041 | .125 | - | .250 | .250 | - | .166 | .125 | .041 | .041 | .166 | - | .083 | .041 | - | .041 | .125 | .416 | .166 |
| Man | 23 | .050 | .050 | .100 | .150 | - | .100 | .100 | .050 | .100 | .250 | .050 | .200 | .150 | .100 | .100 | .100 | .150 | - | .150 | .100 | .150 | .150 | 1 | .250 | .050 | - | .100 | .200 | - | - | .050 | .150 | .150 | .100 | .350 | .150 | - | .250 | .050 | .100 | .100 | .050 | - | .050 | .050 | .200 | - |
| Materials | 24 | .125 | .208 | .125 | .083 | .041 | .083 | .250 | .083 | .416 | .041 | - | .125 | .083 | .208 | .083 | .125 | .166 | .125 | .125 | .041 | .083 | .375 | .208 | 1 | .041 | .041 | .041 | - | .083 | - | - | .083 | .291 | .125 | .250 | .166 | - | - | .041 | - | .083 | .041 | .041 | .083 | .291 | .125 |
| Mathematics | 25 | - | - | - | - | - | .142 | .142 | - | - | - | .428 | .142 | .285 | - | - | - | - | - | .142 | .142 | - | .285 | .142 | .142 | 1 | - | .142 | .142 | - | - | .142 | .285 | - | .142 | .142 | - | - | .142 | - | - | .142 | .285 | .142 | .142 | - | - |
| Measurement | 26 | .125 | - | .125 | - | .250 | - | - | - | .125 | - | .125 | - | .250 | - | - | - | .500 | .125 | .125 | - | .125 | .125 | - | .625 | .125 | - | .125 | - | .25 | - | - | .625 | .125 | - | .125 | .125 | - | - | - | .125 | .125 | .125 | - | - | - | - | .250 |
| Missiles and Rockets | 27 | .333 | - | - | - | - | .333 | .444 | - | - | - | - | .222 | - | - | .333 | - | .222 | - | - | .222 | - | .222 | .222 | .111 | .111 | .111 | 1 | - | .111 | .111 | - | .333 | .333 | - | .444 | .333 | - | - | .111 | .222 | - | .222 | - | - | - | .333 | - |
| Mystery, Myths, & Prob. | 28 | .285 | - | .142 | .142 | - | - | - | - | .285 | - | .142 | .142 | - | .142 | .285 | - | .285 | .142 | - | .142 | .142 | .285 | .571 | - | .142 | - | - | 1 | .142 | - | .142 | .142 | .142 | - | .142 | .285 | - | .142 | .142 | - | .142 | - | - | .142 | .142 | .285 | - |
| Nature | 29 | .111 | .333 | .333 | - | .111 | .111 | - | .222 | .222 | - | - | - | - | .111 | - | .111 | .444 | .111 | .111 | .111 | .222 | .111 | - | .222 | - | .111 | .111 | .111 | 1 | - | .111 | .222 | .222 | .222 | .111 | .222 | - | - | .111 | - | - | - | - | .222 | - | .111 | .111 |
| Navigation | 30 | .571 | - | - | - | .285 | .142 | .285 | - | - | .142 | - | .142 | .142 | - | .142 | - | .285 | .142 | - | .285 | - | .428 | - | - | - | - | .142 | - | - | 1 | - | .142 | .428 | - | .285 | .142 | .142 | - | - | .285 | .142 | .428 | - | - | - | .428 | - |
| Paleontology | 31 | - | - | .800 | .800 | - | - | - | .400 | - | - | - | - | - | - | - | - | .400 | - | .200 | - | .800 | - | .200 | - | - | - | - | .200 | .200 | - | 1 | - | - | - | .200 | - | - | - | - | - | - | - | - | .200 | - | - | - |
| Physical Quantities | 32 | .312 | .062 | .062 | - | .125 | .250 | .125 | - | .062 | .125 | .125 | - | .250 | .187 | .125 | - | .250 | .125 | .187 | .125 | .062 | .375 | .187 | .125 | .062 | .312 | .187 | .062 | .125 | .062 | - | 1 | .437 | .062 | .125 | .125 | - | .062 | .125 | .062 | - | .125 | - | .062 | .062 | .250 | .312 |
| Physics | 33 | .200 | .050 | .050 | .050 | .200 | .150 | .300 | - | .150 | .150 | - | .100 | .300 | .150 | .050 | - | .200 | .250 | .150 | .150 | - | .300 | .150 | .350 | .100 | .050 | .150 | .050 | .100 | .150 | - | .350 | 1 | .050 | .200 | .300 | .050 | .050 | - | .150 | - | .150 | .050 | .050 | - | .100 | .050 |
| Plants | 34 | .091 | .455 | .182 | .091 | - | - | .091 | .364 | .182 | - | - | - | - | - | - | .273 | .182 | .091 | .091 | .091 | .091 | - | .091 | .273 | - | - | - | - | .091 | .091 | - | - | .091 | .091 | 1 | .273 | - | - | - | - | - | - | .182 | .091 | - | .091 |
| Political or Government etc | 35 | .148 | .111 | .037 | .185 | .111 | .074 | .333 | .037 | .074 | .148 | .037 | .222 | .146 | .037 | .074 | - | .555 | .074 | .074 | .037 | .185 | .148 | .269 | .222 | .047 | .037 | .148 | .037 | .074 | .074 | .037 | .074 | .148 | .111 | 1 | .185 | .074 | .037 | .037 | .185 | .037 | .074 | .037 | .074 | .037 | .074 | .074 |
| Power | 36 | .384 | - | - | .077 | - | .077 | .461 | - | .308 | - | - | .308 | .154 | .154 | .461 | - | .231 | .154 | .154 | .154 | - | .231 | .231 | .308 | .077 | - | .231 | .154 | - | .077 | - | .154 | .461 | - | .384 | 1 | - | - | - | .154 | - | .077 | .077 | .077 | - | .231 | - |
| Predictions | 37 | .400 | - | - | - | .200 | .200 | - | - | - | .200 | .200 | - | .600 | - | - | - | - | - | - | .200 | - | .200 | - | - | - | - | - | - | .200 | - | - | .200 | - | .400 | - | .1 | - | .200 | .400 | - | .200 | - | - | - | .600 | .800 |
| Psychology | 38 | .166 | - | .166 | - | - | - | - | - | - | - | .166 | .166 | .166 | - | - | - | - | - | - | .333 | - | .166 | .833 | - | - | - | - | .166 | .166 | - | - | .166 | .166 | - | .166 | - | - | 1 | - | - | .333 | - | - | - | - | .333 | - |
| Research | 39 | .428 | .142 | - | - | - | .285 | - | - | - | - | .285 | - | .142 | .285 | - | .142 | .285 | .142 | .428 | - | - | .571 | .142 | .142 | .142 | .142 | .142 | .142 | - | - | - | .285 | - | - | .142 | - | .142 | - | 1 | .142 | - | - | - | .142 | .285 | .571 |
| Satellites | 40 | .200 | - | - | - | .400 | .300 | .100 | - | - | .200 | - | - | .400 | - | .200 | - | .300 | - | .200 | - | - | - | .200 | - | - | .100 | .200 | - | - | .200 | - | .100 | .300 | - | .500 | .200 | .200 | - | .100 | 1 | - | .200 | - | - | - | .200 | .100 |
| Social Sciences | 41 | .125 | - | - | - | - | - | - | - | - | - | - | .125 | - | - | - | - | .250 | - | - | - | .125 | .250 | .250 | - | .125 | - | .125 | - | .125 | - | - | - | - | .125 | - | - | .250 | - | - | 1 | - | - | - | - | .375 | - |
| Space Travel | 42 | .142 | - | - | - | .428 | .428 | .428 | - | - | - | - | - | .142 | - | .142 | .142 | .142 | .142 | - | .142 | - | .142 | .142 | .142 | .142 | - | .285 | - | .428 | - | - | .285 | .428 | .285 | .142 | .142 | - | - | .285 | - | 1 | - | .142 | .142 | - | .142 | - |
| Teaching Education | 43 | - | - | - | - | - | - | .200 | .200 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | .400 | - | - | - | - | - | - | .200 | - | .200 | .200 | - | - | - | - | - | 1 | - | - | - | - | - | - |
| Time | 44 | .142 | .285 | .142 | .142 | .142 | .142 | .142 | - | - | - | - | - | .285 | .142 | .142 | .142 | .571 | - | - | .142 | .142 | .142 | .285 | .142 | - | - | .142 | .285 | - | - | .142 | .142 | .285 | .285 | .142 | - | - | - | .142 | - | - | - | 1 | .142 | - | - | - |
| Tools | 45 | .166 | .166 | .166 | .166 | - | .333 | .166 | - | - | - | .333 | - | .166 | .166 | - | .166 | - | - | .166 | - | .166 | .500 | .166 | .166 | .166 | - | - | .166 | - | - | .166 | .166 | - | .166 | - | - | - | .166 | - | - | - | .166 | 1 | .166 | .166 | .166 |
| Transportation | 46 | .687 | - | .062 | - | - | .250 | .125 | - | .062 | .062 | .062 | .250 | .062 | .125 | .312 | - | .062 | - | .062 | .312 | - | .625 | .250 | .062 | - | - | .187 | .125 | .062 | .187 | - | .250 | .125 | - | .125 | .187 | .187 | .125 | .125 | .125 | .187 | .062 | - | - | .062 | 1 | .125 |
| Weather | 47 | .666 | .111 | - | - | - | .222 | - | - | - | .111 | .222 | - | .333 | .222 | .111 | - | .222 | .111 | .222 | .111 | .111 | .444 | - | .222 | - | .222 | - | - | .111 | - | - | .555 | .111 | .111 | .222 | - | .444 | - | .444 | .111 | - | - | - | - | .111 | .222 | 1 |

| $I_j$ | | $I_k$ | | $P(I_j, I_k)$ |
|---|---|---|---|---|
| 1 | Aerodynamics and Aviation | 46 | Transportation | 0.65 |
| 2 | Agriculture | 24 | Materials | 0.63 |
| | | 34 | Plants | |
| 3 | Animals (including birds, fish, and reptiles) | 8 | Biology | 0.38 |
| 4 | Archaeology | 21 | History | 0.88 |
| 5 | Astronomy | 33 | Physics | 0.31 |
| | | 40 | Satellites | |
| 6 | Atmosphere | 1 | Aerodynamics and Aviation | 0.31 |
| | | 22 | Machinery | |
| | | 32 | Physical Quantities | |
| | | 46 | Transportation | |
| 7 | Atomic Physics | 35 | Political or government groups or functions | 0.64 |
| 8 | Biology | 3 | Animals | 0.45 |
| 9 | Chemistry | 24 | Materials | 0.67 |
| 10 | Communications | 23 | Man | 0.50 |
| 11 | Computers | 13 | Electronics | 0.71 |
| 12 | Defense and Warfare | 35 | Political or government groups or functions | 0.55 |
| 13 | Electronics | 22 | Machinery | 0.43 |
| | | 33 | Physics | |
| 14 | Engineering | 22 | Machinery | 0.64 |
| 15 | Engines | 1 | Aerodynamics and Aviation | 0.75 |
| | | 36 | Power | |
| 16 | Food | 2 | Agriculture | 0.43 |
| | | 24 | Materials | |
| | | 34 | Plants | |
| 17 | Geography | 35 | Political or government groups or functions | 0.71 |
| 18 | Geology | 33 | Physics | 0.63 |
| 19 | Geophysics | 13 | Electronics | 0.40 |
| | | 17 | Geography | |
| 20 | Health and Safety | 1 | Aerodynamics and Aviation | 0.38 |
| | | 46 | Transportation | |
| 21 | History | 4 | Archaeology | 0.70 |
| | | 17 | Geography | |
| 22 | Machinery | 1 | Aerodynamics and Aviation | 0.42 |
| | | 46 | Transportation | |
| 23 | Man | 35 | Political or government groups or functions | 0.35 |
| 24 | Materials | 9 | Chemistry | 0.42 |
| 25 | Mathematics | 11 | Computers | 0.43 |
| 26 | Measurement | 32 | Physical Quantities | 0.63 |
| 27 | Missiles and Rockets | 7 | Atomic Physics | 0.44 |
| | | 35 | Political or government groups or functions | |
| 28 | Mystery, Myths and Problems | 23 | Man | 0.57 |
| 29 | Nature | 17 | Geography | 0.44 |
| 30 | Navigation | 1 | Aerodynamics and Aviation | 0.57 |
| 31 | Paleontology | 3 | Animals | 0.80 |
| | | 4 | Archaeology | |
| | | 21 | History | |
| 32 | Physical Quantities | 33 | Physics | 0.44 |
| 33 | Physics | 24 | Materials | 0.35 |
| | | 32 | Physical Quantities | |
| 34 | Plants | 2 | Agriculture | 0.45 |
| 35 | Political or government groups or functions | 17 | Geography | 0.56 |
| 36 | Power | 7 | Atomic Physics | 0.46 |
| | | 15 | Engines | |
| | | 33 | Physics | |
| 37 | Predictions | 47 | Weather | 0.80 |
| 38 | Psychology | 23 | Man | 0.83 |
| 39 | Research | 22 | Machinery | 0.57 |
| | | 47 | Weather | |
| 40 | Satellites | 35 | Political or government groups or functions | 0.50 |
| 41 | Social Sciences | 46 | Transportation | 0.38 |
| 42 | Space Travel | 5 | Astronomy | 0.43 |
| | | 6 | Atmosphere | |
| | | 7 | Atomic Physics | |
| | | 30 | Navigation | |
| | | 33 | Physics | |
| 43 | Teaching - Education | 25 | Mathematics | 0.40 |
| 44 | Time | 17 | Geography | 0.57 |
| 45 | Tools | 22 | Machinery | 0.50 |
| 46 | Transportation | 1 | Aerodynamics and Aviation | 0.69 |
| 47 | Weather | 1 | Aerodynamics and Aviation | 0.67 |

LIST OF MOST HIGHLY CORRELATED INDEX TERMS

(Forward Conditionals)

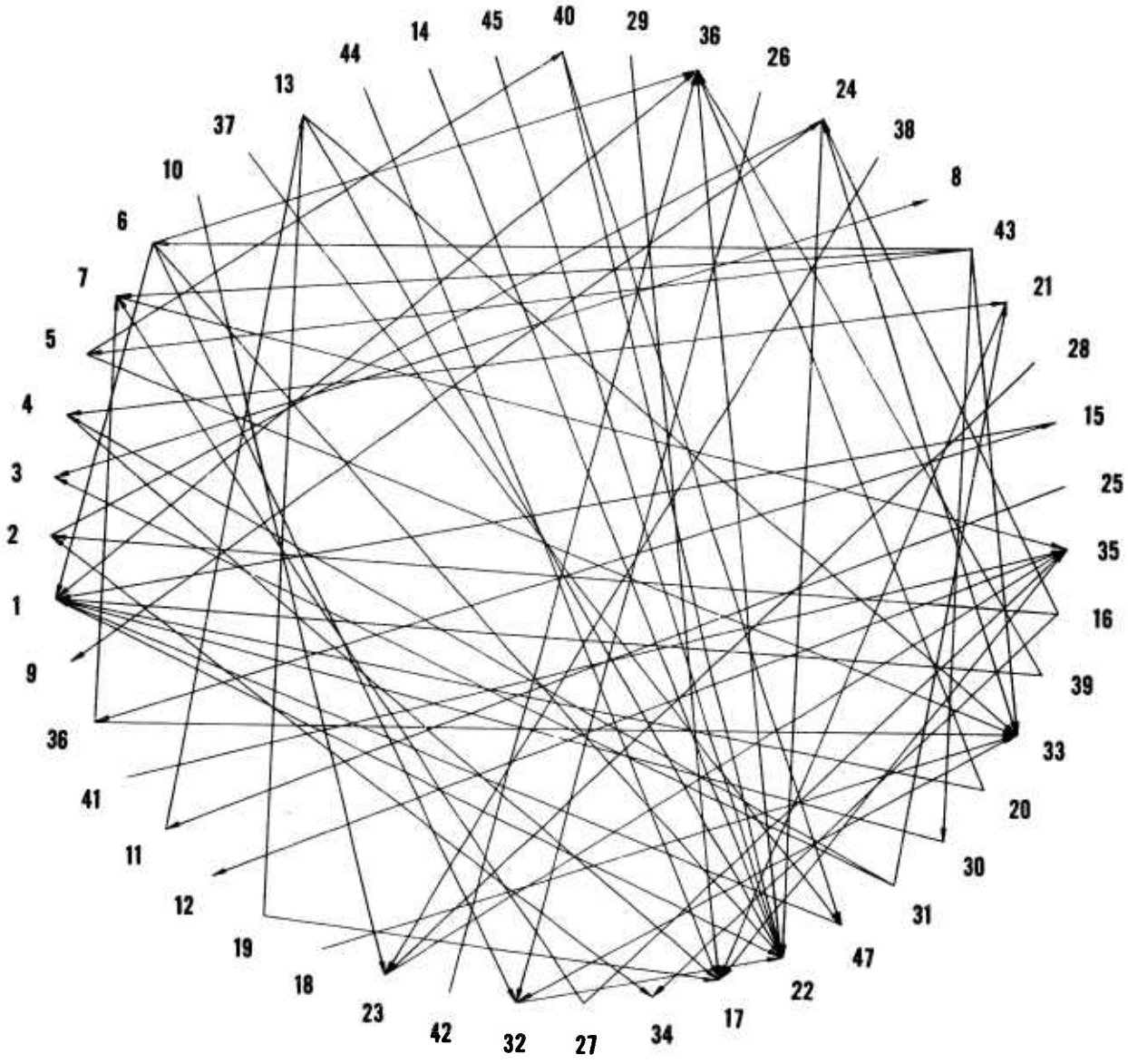
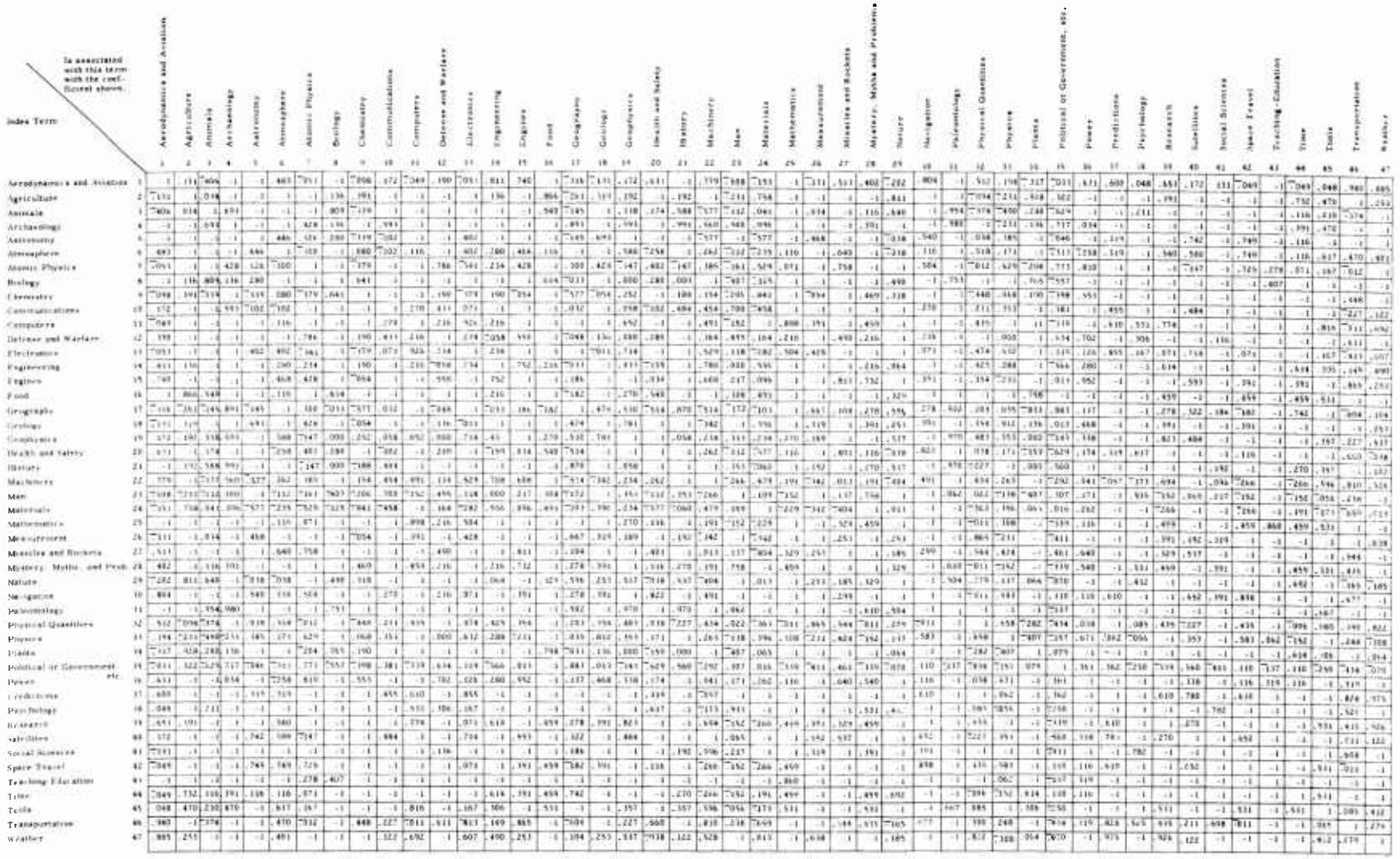



DIAGRAM 5. A MAP OF INDEX SPACE

| $I_j$ | | $I_k$ | | $P(I_k, I_j)$ |
|---|---|---|---|---|
| 1 | Aerodynamics and Aviation | 15 | Engines | 0.75 |
| 2 | Agriculture | 34 | Plants | 0.46 |
| 3 | Animals (including birds, fish, and reptiles) | 31 | Biology | 0.80 |
| 4 | Archaeology | 31 | Biology | 0.80 |
| 5 | Astronomy | 42 | Space Travel | 0.43 |
| 6 | Atmosphere | 42 | Space Travel | 0.43 |
| 7 | Atomic Physics | 36 | Power | 0.46 |
| 8 | Biology | 31 | Paleontology | 0.40 |
| 9 | Chemistry | 24 | Materials | 0.42 |
| 10 | Communications | 13 | Electronics | 0.29 |
| 11 | Computers | 25 | Mathematics | 0.43 |
| 12 | Defense and Warfare | 7 | Atomic Physics | 0.36 |
| 13 | Electronics | 11 | Computers | 0.71 |
| 14 | Engineering | 15 | Engines | 0.38 |
| 15 | Engines | 36 | Power | 0.46 |
| 16 | Food | 2 | Agriculture | 0.38 |
| 17 | Geography | 4 | Archaeology | 0.75 |
| 18 | Geology | 19 | Geophysics | 0.30 |
| 19 | Geophysics | 39 | Research | 0.43 |
| 20 | Health and Safety | 38 | Psychology | 0.33 |
| 21 | History | 4 | Archaeology | 0.88 |
| 22 | Machinery | 14 | Engineering | 0.64 |
| 23 | Man | 38 | Psychology | 0.83 |
| 24 | Materials | 9 | Chemistry | 0.67 |
| 25 | Mathematics | 11 | Computers | 0.43 |
| 26 | Measurement | 32 | Physical Quantities | 0.31 |
| 27 | Missiles and Rockets | 15 | Engines | 0.38 |
| 28 | Mystery, Myths and Problems | 15 | Engines | 0.25 |
| 29 | Nature | 2 | Agriculture | 0.38 |
| 30 | Navigation | 42 | Space Travel | 0.43 |
| 31 | Paleontology | 4 | Archaeology | 0.50 |
| 32 | Physical Quantities | 26 | Measurement | 0.62 |
| 33 | Physics | 18 | Geology | 0.62 |
| 34 | Plants | 2 | Agriculture | 0.62 |
| 35 | Political or government groups or functions | 17 | Geography | 0.71 |
| 36 | Power | 15 | Engines | 0.75 |
| 37 | Predictions | 47 | Weather | 0.44 |
| 38 | Psychology | 23 | Man | 0.25 |
| | | 41 | Social Sciences | |
| 39 | Research | 47 | Weather | 0.44 |
| 40 | Satellites | 37 | Predictions | 0.40 |
| 41 | Social Sciences | 38 | Psychology | 0.33 |
| 42 | Space Travel | 30 | Navigation | 0.43 |
| 43 | Teaching - Education | 25 | Mathematics | 0.28 |
| 44 | Time | 2 | Agriculture | 0.25 |
| 45 | Tools | 11 | Electronics | 0.28 |
| 46 | Transportation | 1 | Aerodynamics and Aviation | 0.65 |
| 47 | Weather | 37 | Predictions | 0.80 |

LIST OF MOST HIGHLY CORRELATED INDEX TERMS

(Inverse Conditionals)




Coefficients of Association

| $I_j$ | | $I_k$ | | $Q(I_j, I_k)$ |
|---|---|---|---|---|
| 1 | Aerodynamics and Aviation | 46 | Transportation | 0.940 |
| 2 | Agriculture | 34 | Plants | 0.928 |
| 3 | Animals (including birds, fish, and reptiles) | 31 | Paleontology | 0.954 |
| 4 | Archaeology | 21 | History | 0.991 |
| 5 | Astronomy | 42 | Space Travel | 0.749 |
| 6 | Atmosphere | 42 | Space Travel | 0.749 |
| 7 | Atomic Physics | 36 | Power | 0.810 |
| 8 | Biology | 3 | Animals | 0.809 |
| 9 | Chemistry | 24 | Materials | 0.841 |
| 10 | Communications | 23 | Man | 0.700 |
| 11 | Computers | 13 | Electronics | 0.926 |
| 12 | Defense and Warfare | 7 | Atomic Physics | 0.786 |
| 13 | Electronics | 11 | Computers | 0.926 |
| 14 | Engineering | 1 | Aerodynamics and Aviation | 0.811 |
| 15 | Engines | 36 | Power | 0.952 |
| 16 | Food | 2 | Agriculture | 0.866 |
| 17 | Geography | 4 | Archaeology | 0.891 |
| 18 | Geology | 19 | Geophysics | 0.781 |
| 19 | Geophysics | 31 | Paleontology | 0.970 |
| 20 | Health and Safety | 30 | Navigation | 0.822 |
| 21 | History | 4 | Archaeology | 0.991 |
| 22 | Machinery | 14 | Engineering | 0.788 |
| 23 | Man | 38 | Psychology | 0.935 |
| 24 | Materials | 9 | Chemistry | 0.841 |
| 25 | Mathematics | 11 | Computers | 0.898 |
| 26 | Measurement | 32 | Physical Quantities | 0.865 |
| 27 | Missiles and Rockets | 15 | Engines | 0.811 |
| 28 | Mystery, Myths and Problems | 23 | Man | 0.758 |
| 29 | Nature | 2 | Agriculture | 0.811 |
| 30 | Navigation | 42 | Space Travel | 0.898 |
| 31 | Paleontology | 4 | Archaeology | 0.980 |
| 32 | Physical Quantities | 26 | Measurement | 0.865 |
| 33 | Physics | 18 | Geology | 0.812 |
| 34 | Plants | 2 | Agriculture | 0.928 |
| 35 | Political or Government groups or functions | 17 | Geography | 0.883 |
| 36 | Power | 15 | Engines | 0.952 |
| 37 | Predictions | 47 | Weather | 0.975 |
| 38 | Psychology | 23 | Man | 0.935 |
| 39 | Research | 47 | Weather | 0.926 |
| 40 | Satellites | 37 | Predictions | 0.780 |
| 41 | Social Sciences | 38 | Psychology | 0.782 |
| 42 | Space Travel | 30 | Navigation | 0.898 |
| 43 | Teaching - Education | 25 | Mathematics | 0.860 |
| 44 | Time | 17 | Geography | 0.742 |
| 45 | Tools | 11 | Computers | 0.816 |
| 46 | Transportation | 1 | Aerodynamics and Aviation | 0.940 |
| 47 | Weather | 37 | Predictions | 0.975 |

LIST OF MOST HIGHLY CORRELATED INDEX TERMS

(Coefficient of Association)

as was reasonable. They were told the meaning of the logical connections (conjunction and disjunction) and they were allowed to use these connectives as desired. Having formulated a library request for each of the eight questions they were asked to repeat the process, the second time weighting (as desired) the index tags used in the requests. We took their library requests searched, computed, ranked the retrieved documents and evaluated the results with the aid of the individual who originally made the requests. The results are described in the following sections.

## 2. THE MEASURE OF RELEVANCE

### 2.1 Initial Remarks

In this section we describe our attempts to evaluate empirically the goodness of our function which gives a relevance number. That is to say, in Part II, 1, we explicated the notion of probable relevance via Bayes' Theorem and in the present section we shall describe and interpret the results of a test to evaluate empirically the goodness of our relevance function. Thus, again, the problem which we shall now consider is: "How well does our function perform in ranking documents according to relevance"? Only in section 3 shall we examine the other major question; viz., "How good is the selection process"?

### 2.2 Some Clarification

Let us briefly review the notion of relevance. Exactly what is it we wish to measure? First, we must be careful to distinguish a user's information need N from his request formulation R. In our experiments we can think of N as the particular information item desired, R as the formulation of this item in the request language. Thus, in the example question given in Table 12, p83, we have

N: Turbojet engines for commercial air travel,
R: Transportation and (Aerodynamics-Aviation or Engines).

In a real library system N will never be known, only its description R in a rather artifical language. The library indexing system only relates documents to this request language. But we want to relate documents to information need. We see then that a bridge between request language and information need is through statistical data relating library users with the utility they derive from documents. Such statistical data is given by the a priori probability distribution. This is shown by the theoretical development which states that the probability of a document satisfying the request, i.e., the probability of a document giving the desired information item N, is proportional to the product of the a priori probability $(P(A, D_i))$ and the value $(\omega_i(R))$ of the extended weight function for the request describing that need. In a sense then this quantity is a measure

of the degree of relevance of a document for the information need of the requestor. We say "in a sense" because of its unavoidable probabilistic nature. It is a probabilistic estimate of the relevance of a document for the information need of the requestor. With this qualification in mind we call this quantity "relevance to information need" or "probable relevance". We have:

$$(\text{relevance to information need}) \sim P(A, D_i) \cdot \omega_i(R) \qquad (1)$$

## 2.3 An Experimental Result

If we examine again Table 1 (p. 21) which gives the various interpretations of the request language we see that it might be fruitful to look for a quantitative measure of relevance with respect to the logical meaning of the request; i.e., a measure of "relevance to request" as opposed to "relevance to information need". We conjecture that our computational procedure for computing the values of the extended weight function is such a measure. We will present the supporting data in subsequent sections, but in anticipation of this we state the result: Bayes' Theorem plus experimental data implies:

$$(\text{relevance to information need}) \sim P(A, D_i) \cdot (\text{relevance to request formulation}). \qquad (2)$$

## 2.4 The Result Predicted by Bayes' Theorem

The content of Bayes' Theorem (formula (1)) can be illustrated by the following hypothetical experiment: Consider a document in the experimental library. It consists of many information items. Select one of these. Let a library user formulate this item in the library request language. Let the library system now operate on the request, producing a collection of documents. (If the library indexing system is adequate and the formulation of the request is accurate, the original document from which the information item was derived should appear in this retrieved collection.) We now ask the library user to prepare a list:

$L_1$: the retrieved documents ranked according to relevance to the information item.

We ask another person to prepare a second list:

$L_2$: the retrieved documents ranked according to relevance to the request formulation.

To facilitate the processing of this comparative data we ask that the documents be classified into five categories: I, Very Relevant; II, Relevant; III, Somewhat Relevant; IV, Only Slightly Relevant; V, Irrelevant.

Suppose now we simulate an a priori probability distribution, and repeat the above experiment many times, each time selecting a document by using the simulated distribution. For each request we obtain lists $L_1$ and $L_2$ and a third list $L_3$:

$L_3$: the retrieved documents ranked according to the magnitude

$$P(A, D_i) \cdot \omega_i(R).$$

Bayes' Theorem now tells us what we may expect to find; namely, that the list $L_3$ will agree with list $L_1$ in the long run. More precisely, for each request R there are many information items (or needs) that would be formulated by R, one for each requestor who uses R. If, for each list $L_1$ that originated with these requestors, we computed the mean relevance evaluation for each document in $L_1$ by using the category numbers I, II, III, IV, V, then the resulting ranking should agree with list $L_3$.

## 2.5 The Experimental Design

The result predicted above is difficult to test empirically because it would require such a large sample, but, an experiment designed on a much smaller scale can give us some valuable information. Such an experiment is the one described in section 1.5 and the following sections. Since

it is designed primarily to test both the basic selection process and the search strategy by elaborating the request, as well as the computational schema for $\omega_i(R)$, a flat a priori probability distribution is assumed, i.e., all $P(A, D_i)$ are taken to be equal. The significance of this for the probable relevance concept is clear by looking at formula (1):

$$\text{(relevance to information need)} \sim \omega_i(R). \tag{3}$$

The interpretation of this by the phase "in the long run" still holds however. That is, we would not expect a single list of type $L_1$ to compare with its corresponding list of type $L_3$ (this last being the ranking of documents by the values $\omega_i(R)$ in the case of equal $P(A, D_i)$). This can be seen by noting that as the information need becomes more specific the evaluations in a list of type $L_1$ would tend to split into the two classifications of Very Relevant or Irrelevant, but the ranking by the values of $\omega_i(R)$ always varies gradually. On the other hand a list of type $L_2$ might conceivably be expected to agree with the list $L_3$ in a single case. This is the content of the experimental result stated in 2.3. To bear this out we selected eight of the 40 requests and obtained for each of these requests a list of type $L_2$. In addition we had control lists of types $L_1$ and $L_2$ prepared for these same requests (i.e., each evaluation done twice by different persons). The lists of type $L_2$ had, as expected, a more even distribution of documents throughout the five categories I (Very Relevant) to V (Irrelevant).

The processing of this evaluation data was accomplished on the following lines: We saw what comparative data was reflected in the scale of values $\omega_i(R)$ and compared this with the evaluations of the $L_2$ lists. The details of the experimental data and its analysis are presented in the following sections (2.6-2.8).

## 2.6 The Hypotheses to be Tested

We can formulate our goal as that of attempting to confirm that the value $\omega_i(R)$ that we compute for each document selected by a given request is,

in fact, a measure of relevance with respect to the request formulation. If our basic notion is correct it implies the following hypothesis which we call $H_1$.

$H_1$: if a document is relevant to a request, then a high number $\omega_i(R)$ will be derived for it.

How to verify, confirm, test this hypothesis empirically? We did the following: A number of documents from our experimental library were selected at random and, for each document, a question was formulated which could be answered by reading the corresponding document. Several persons who acted as test subjects were briefed as to the nature of the library, the indexing system, etc., given a set of questions and asked to formulate a library request for information on the basis of which, hopefully, relevant documents would be retrieved (so as to answer the question). Given the library requests that these test subjects formulated we proceeded to search and select the accession numbers of those documents satisfying the logic of the request. For each request a list of documents (i.e., a list of the corresponding accession numbers) was generated and the documents in the lists were ranked according to the number $\omega_i(R)$ that was computed for each. We then examined each list to determine whether or not the so-called "answer" document was on the list and if it was, we recorded its relative position on the list. We made the (natural) assumption that the answer document (i.e., the document on the basis of which the question was formulated) would be relevant to the request. We then determined the number of times that the correct answer document was retrieved associated with a high number $\omega_i(R)$. The results can be summarized as follows: Forty library requests were made and in 27 cases the answer document was retrieved. The number of documents on the output lists ranged from a minimum of 1 (in four cases) to a maximum of 41. In the majority of the 23 cases which contained more than a single document, the answer document appeared towards the top of the list.

The results showed that if the answer document was on a list, then it was computed to have a high number $\omega_i(R)$ in most of the cases. This evidence thus supports the hypothesis $H_1$, which asserts that if a document is relevant a high value $\omega_i(R)$ will be computed for it. However, it was not always the case that the answer document was computed to have the highest number; i.e., there were documents other than the answer document for which a high number was derived. Thus, the question arises: "If a document has a high number $\omega_i(R)$, is it relevant to the request"? This represents the converse of the original hypothesis $H_1$. We shall form this as an hypothesis and call it $H_2$.

$H_2$: if a document has a high number $\omega_i(R)$, then it is relevant to the corresponding request.

If we can confirm $H_2$ as well as $H_1$ we will have, in fact, confirmed an hypothesis $H^*$ which is stronger than each.

$H^*$: the methods of Probabilistic Indexing will derive a high number $\omega_i(R)$ for an arbitrary document <u>if</u> and <u>only if</u> the document in question is relevant to the request.

In order to determine if there were relevant documents, other than the answer document on a list we had to have evaluation data of the type described in 2.5. We obtained a sample of this information from the test subjects in the following way: Four of the five test subjects were given the actual documents corresponding to the retrieval lists and they were asked to read each document and decide whether they considered it to be <u>Very Relevant</u>, <u>Relevant</u>, <u>Somewhat Relevant</u>, <u>Only Slightly Relevant</u>, <u>Irrelevant</u>. Thus for each document retrieved they would judge to which of these five categories it belonged and we, in turn, compared their judgments with the numbers $\omega_i(R)$ which we had computed for each document. A fifth person prepared control lists, i.e., evaluations for the same requests.

In order to facilitate the comparison we standardized the values $\omega_i(R)$; i.e., we multiplied each value by the reciprocal of the highest value to force the numbers on each list to vary from 1 to 0 - 0 being the value assigned to unretrieved documents. We also divided the numbers into three categories: high (value equal to or greater than 0.75), medium (value between 0.75 and 0.25) and low (value equal to or less than 0.25). The results show quite definitely that if a document has a high number $\omega_i(R)$ that document was judged by the evaluator as Very Relevant or Relevant, in most cases. Conversely, if the number $\omega_i(R)$ was low the evaluators rated the corresponding document as either Only Slightly Relevant or Irrelevant in most cases.

Thus the data supports the following: "If a document is relevant to a request, then there is a strong probability that the document will have a a high number $\omega_i(R)$ computed for it." Furthermore, the data supports the converse: viz., "If a document is computed to have a high number $\omega_i(R)$, there is a strong probability that it is relevant to the request". Thus the data supports both $H_1$ and $H_2$ and taken jointly we see that the data does support and confirm the stronger hypothesis $H^*$, viz., a high number $\omega_i(R)$ will be derived if and only if the document in question is relevant to the request. The details of the analysis are presented in the following section.

## 2.7 Analysis of the Data

The eight lists involved in the evaluation with respect to request relevance had a sum total of 69 documents. First we examine how the values $\omega_i(R)$ associated with these documents were distributed among the five categories. Computing the average value and the variance in each of the five categories, we obtained the following results.

| DOCUMENT RATING | | MEAN | VARIANCE |
|---|---|---|---|
| I. | Very Relevant | 0.81 | 0.043 |
| II. | Relevant | 0.72 | 0.053 |
| III. | Somewhat Relevant | 0.54 | 0.043 |
| IV. | Only Slightly Relevant | 0.40 | 0.110 |
| V. | Irrelevant | 0.18 | 0.013 |

Thus we see that the values of the numbers that we computed decrease, on the average, as we go from Category I (Very Relevant) to Category V (Irrelevant).

Although this result tends to confirm our hypotheses $H_1$, $H_2$, $H^*$ we prefer to look deeper into the situation. Let us denote the class of all documents with numbers $\omega_i(R)$ greater or equal to 0.75 by "High"; those with numbers less than or equal to 0.25 by "Low". Let us also call categories I or II simply "Relevant" and category V, as before, "Irrelevant". Note that Relevant and Irrelevant are not negations of each other since we have the intermediate categories III and IV consisting of documents neither totally Relevant nor Irrelevant. Now the hypotheses $H_1$, $H_2$, $H^*$ say two things:

(1) Relevant is equivalent to High,
(2) Irrelevant is equivalent to Low,

and imply two weaker statements:

(3) Relevant implies not-Low,
(4) Irrelevant implies not-High.

The statistical confirmation of these statements can be accomplished by using the theory of the coefficient of association between predicates as outlined in section 2.6. That is to say, each of the statements above calls for a study of a matrix of the kind defined on p. 42; i.e., a sorting of the 69 documents according to the properties:

(1) Relevant and High,

(2) Irrelevant and Low,

(3) Relevant and Low,

(4) Irrelevant and High.

We would expect to find the Q-values in (1) and (2) to be near +1 (maximum positive association), the Q-values in (3) and (4) near -1 (maximum negative association). These values are in fact:

$$Q(\text{Relevant, High}) = +0.70,$$
$$Q(\text{Irrelevant, Low}) = +0.90,$$
$$Q(\text{Relevant, Low}) = -0.92,$$
$$Q(\text{Irrelevant, High}) = -1.00.$$

Since these values are fairly sensitive we introduce a control on the study by assuming that these predicates are statistically independent, then computing the probabilities of the Q-values having been as close or closer to the anticipated values by chance. For the four distributions we calculate these control probabilities to be 0.041, 0.006, 0.010, 0.059, respectively. [1]

## 2.8 A Note on other Data

We have still to consider the weighted request. Recall that we have two types of inputs to consider; viz., the conventional request (an affirmative Boolean function of the index terms) and the weighted request. These lead to two different output lists. On page 83 we include a typical data tabulation sheet for one of the questions that we used in our experiments. The conventional request in this case was expressed formally by the expression

$$I_{46} \cdot (I_1 \vee I_{15})$$

[1] That is to say, if Relevant and High are independent then the probability of their Q-value having the property $0.70 \leqslant Q \leqslant 1.00$ is 0.041. Similarly for the other classifications.

Question: Turbojet Engines for Commercial Air Travel.

Answer Document: 92

Request: Transportation and (Aerodynamics-Aviation or Engines)

Weighted Request: (.8) Transportation and [(.3) (Aero-Aviation) or (.9) Engines]

| B | P | | W | | P' | | W' | | Evaluation with respect to request | |
|---|---|---|---|---|---|---|---|---|---|---|
| Accession Number | Acc. No. | Rel. No. | Acc. No. | Rel. No. | Acc. No. | Rel. No. | Acc. No. | Rel. No. | Acc. No. | Evaluation |
| 7 | 36 | 1.000 | 36 | .779 | 36 | .0211 | 36 | .0164 | 92 | Very Relevant |
| 8 | 92 | .875 | 92 | .708 | 92 | .0200 | 92 | .0161 | 36 | Relevant |
| 33 | 104 | .546 | 7 | .408 | 104 | .0076 | 7 | .0060 | 104 | Relevant |
| 36 | 8 | .437 | 108 | .270 | 7 | .0050 | 8 | .0037 | 8 | Somewhat Relevant |
| 61 | 61 | .390 | 8 | .254 | 108 | .0049 | 108 | .0035 | 61 | Somewhat Relevant |
| 73 | 108 | .375 | 90 | .225 | 61 | .0032 | 73 | .0028 | 7 | Only Slightly Relevant |
| 77 | 7 | .339 | 73 | .225 | 77 | .0017 | 90 | .0026 | 77 | Only Slightly Relevant |
| 90 | 33 | .125 | 104 | .131 | 8 | .0014 | 104 | .0018 | 101 | Only Slightly Relevant |
| 92 | 101 | .109 | 61 | .075 | 33 | .0014 | 61 | .0006 | 33 | Irrelevant |
| 101 | 77 | .109 | 33 | .030 | 101 | .0013 | 77 | .0004 | 73 | Irrelevant |
| 104 | 90 | .093 | 101 | .026 | 73 | .0012 | 33 | .0003 | 90 | Irrelevant |
| 108 | 73 | .093 | 77 | .026 | 90 | .0011 | 101 | .0003 | 108 | Irrelevant |

TABLE 12.

A DATA TABULATION SHEET

and the computation by Bayes' schema resulted in the listing shown in column P. The corresponding weighted request was expressed formally as

$$(0.8)I_{46} \cdot \left[(0.3)I_1 \vee (0.9)I_{15}\right]$$

and the results are shown in column W. For completeness we did calculations using a simulated non-linear a priori probability distribution. The resulting relevance numbers are shown in the two columns labelled P' and W'. The answer document for this particular question was document 92 and, quite by chance, it appears in the second position on each of these four lists.

The results for the case of the weighted request (flat a priori probability distribution assumed) do confirm the basic thesis which asserts that the number $\omega_i(R)$ is, in fact, a measure of relevance with respect to request; however, the data are not as confirmatory as for the case when the request is an affirmative Boolean function. The reason for this is that the evaluations of document relevance were oriented toward the unweighted request.[1] However a consideration of the variation of the mean value of $\omega_i(R)$ is still of interest. Analogous to the table on p. 81 we have:

| | DOCUMENT RATING | MEAN | VARIANCE |
|---|---|---|---|
| I. | Very Relevant | 0.87 | 0.031 |
| II. | Relevant | 0.53 | 0.095 |
| III. | Somewhat Relevant | 0.39 | 0.076 |
| IV. | Only Slightly Relevant | 0.45 | 0.123 |
| V. | Irrelevant | 0.33 | 0.073 |

[1] For example: An evaluation of document relevance with respect to the request "Psychology or Teaching" would give quite different results than when evaluated with respect to the request "(.7) Psychology or (.3) Teaching".

## 3. ELABORATION OF THE SELECTION PROCESS

### 3.1 Initial Remarks

The basic aim behind Probabilistic Indexing has been the obvious one; viz., to improve retrieval effectiveness. The fundamental notion has been to introduce arithmetic (as opposed to logic alone) into the problem of indexing and thereby pave the way for the use of mathematical operations so as to compute probable relevance. Thus the fundamental notion which acts as a wedge to drive an opening into the basic problem of retrieval effectiveness is that of the relevance number (as explicated in terms of the theorem of Bayes). The relevance number, as we have seen, provides a means of ranking documents according to their probable relevance. However, the solution to the problem of retrieval effectiveness involves more than ranking by relevance--it involves the proper selection of those documents which are to be ranked. Before we describe the results of the experiments that were conducted to test our methods for improving the selection process, let us take one more look at the relevance number as a filter to eliminate low relevance documents. In particular, let us consider the usefulness of the relevance number on unelaborated requests.

In our experiments, 40 different library requests were made and a total of 379 documents were retrieved (using the basic process of selecting those documents whose tags are logically compatible with the logic and tags of the request). Let us compare the results of probabilistic searching and so-called "binary" or conventional searching. We can do this by assuming that all the tags which are assigned to documents with a non-zero weight are, in fact, assigned to the corresponding documents in the conventional system. Thus when the basic selection process is the same (viz., the unelaborated logical matching process), the same documents will be retrieved in both cases; however, in the conventional system the retrieved documents are not ranked by any criteria of relevance. For each of the retrieval lists if n documents have been retrieved and the answer document is present, then using the conventional search technique the requestor must read, on the average $\frac{n+1}{2}$ documents. If the answer

document is not present, then all of the retrieved documents must be read (in order to determine that no relevant information was retrieved). These considerations (inadequate though they be, since they presuppose that only an answer document produces a satisfactory search result) give us a criterion with which to compare the probabilistic and binary searches. This criterion is the total number of documents that would have to be read for all 40 searches in order to find the answer documents. The results are as follows:

| Type of Search | Total Number of Documents Retrieved | Total Number of Documents that would have to be Read |
|---|---|---|
| Binary | 379 | 235 |
| Probabilistic [1] | 379 | 181 |

Thus we see that a conventional system would require the user to read approximately 30 percent more retrieved documents to obtain the same number of answer documents. These two different searches, each using the basic selection process, produced 27 answer documents out of a possible 40. (Note that the binary search as defined above is more extensive than might be expected in the sense that we have used all the tags with non-zero weights as binary tags. In an actual conventional system those tags with a low weight would probably not be coordinated with documents. That is to say, the use of weighted tags encourages more tags to be applied to a given document than would be the case if weights were not allowed. In a previous study where documents were indexed independently by two different indexers, one using probabilistic indexing, the other using binary indexing (i.e., either a tag holds for a document or it does not), it was found that 70 percent more answer documents were retrieved in the probabilistic search and only 32 percent more documents had to be read.)

[1] A flat a priori probability distribution was used.

The above comparison presupposes that the user is looking for some specific information (viz., the answer document) and that he knows when he has found it. It might be more realistic to make no such assumption; therefore, let us consider the following comparison. Given a request for information a probabilistic search is made but, beforehand, we tell the user to read only those documents which have a computed relevance number greater than 0.5. That is to say, "before the facts" we give the requestors a guide to use in reading the 40 lists presented to them. It turns out that of the 379 documents in the 40 lists there are only 225 which have a relevance number greater than 0.5. Furthermore, it turns out that if the users had adopted the strategy of reading only those retrieved documents which have relevance numbers greater than 0.5, then they would have found 25 of the 27 answer documents.[1] Now compare this with the case of conventional retrieval where the users would have to read all of the 379 retrieved documents (since there is no way to distinguish between any two documents in the same list). In this latter case the users, of course, would find all 27 answer documents, but again at the "cost" of reading all 379 documents. Thus we see that a conventional system would require that users read 68.5 percent more documents than for the probabilistic system and they would gain only 7.4 percent in increased number of answer documents.

These considerations indicate that the relevance number can be used to filter out irrelevant material. That is to say, if we use the relevance number associated with documents to separate the relevant from the irrelevant, we are providing the user with a valuable tool.

### 3.2 The Automatic Elaboration

We have described two methods for automatically elaborating upon the selection process which is involved in information searching. One method

[1] In one of the two remaining cases the relevance number of the answer document was just under 0.5 and in the other case the answer document had a rather low number but it was third in a list of only three.

establishes a measure of closeness in document space and the other method involves measures of closeness in request space. We shall not consider the former since, as yet no experimental tests have been completed which would enable us to evaluate the notions of distance as described in Part II, 2.7. We measure closeness in request space by determining certain statistical relationships between the index terms of a request and other terms. Specifically, we have described three different statistical relationships, viz., forward conditional probabilities, inverse conditional probabilities, and coefficients of association. We now raise the questions: "How good are the proposed statistical measures of closeness in elaborating upon a request?" and "Which of the three measures that have been discussed is the best?" Again, in the case of the automatically elaborated request we generate the new request R' given the initial request R by formulating the following type of disjunction for each tag in R:

$$\text{if } R = I_j, \text{ then } R' = I_j \vee(\alpha)\, I_j'$$

where $\alpha$ is the measure of closeness between $I_j$ and $I_j'$ and $I_j'$ is the term that gives maximum $\alpha$ with respect to $I_j$. We would like to be able to establish the following:

1. That the elaborated request catches relevant documents which are not selected by the original (unelaborated) request.
2. That, although the elaborated request catches more documents, the relevance number can be used as a guide for eliminating the ones with low probable relevance.

### 3.3 Some Testing (Evaluation) Problems

Since we are really interested in the over-all retrieval effectiveness of the selection process, we would like to know how many of the relevant documents in the entire library have been caught by the elaborated requests. In order to determine this it would be necessary for us to present to the requestor the entire library so that he, in turn, could judge

which relevant documents, if any, were not retrieved. That is to say, in order that a user properly judge whether or not he did, in fact, receive all relevant documents as the result of a search, he would have to be familiar with the entire contents of the library. Because of this difficulty, we see that such an evaluation would be impractical to conduct. We must, therefore, lower our sights and look for a substitute type of evaluation. The substitute that we have adopted consists in, again, using the answer documents as a measure of retrieval effectiveness. That is to say, since we know that the answer documents are relevant, we can automatically elaborate upon those original requests which did not catch the answer document in order to see whether the elaborated request succeeds in retrieving it. Such a test would allow us to establish some measure of the retrieval effectiveness of the automatic elaboration procedures. We can compare the total number of documents for the elaborated requests with what would be the case for the unelaborated request. This we have done and the results are discussed in the following section.

### 3.4 Some Results

Of the 40 requests that were made the answer document was retrieved in 27 cases and it was not retrieved in 13 cases. We conducted three different types of elaborated requests for each of the 40 cases. The results are as follows:

1. Using the method of request elaboration via forward conditional probabilities between index tags, we retrieved the correct answer document in 32 cases out of the 40.
2. Elaborating the requests via the inverse conditional probability heuristic we retrieved the correct document in 33 of the 40 cases.
3. Using the coefficient of association to obtain the elaborated request we obtained success in 33 cases of the 40.

Thus we see that the automatic elaboration of a request does, in fact, catch relevant documents that were not retrieved by the original request.

We now raise the question: "Because of the small size of the library and the large percent of the total library that is selected by the elaborated request, are the above results statistically significant?" That is to say, what is the probability of doing as well or better just by selecting at random, for each of the 13 requests for which the answer document was not originally retrieved, a sample of size equal to that given by the elaborated requests. We have made the corresponding calculations and it turns out that probability of doing as well or better by chance is less than 0.034 for both the forward and inverse conditional probability elaborations and less than 0.001 for the coefficient of association search. Thus the above results are indeed statistically significant.

Could the number of answer documents have been improved; i.e., could 40 out of 40 answer documents have been retrieved. We looked at the seven cases for which the answer document was not retrieved when elaborating via the coefficient of association and in three cases the indexing was at fault. That is to say in three of the seven cases the answer document was poorly indexed (a fact of life that must be faced by all libraries). In one case the request formulation was very poor and no reasonable elaboration would help. In one case the answer document was caught by a different heuristic (viz., the forward conditional) and in the remaining two cases, again, the requests suffered by being poorly formulated.

Now consider the fact that, although the automatic elaboration of a request does catch relevant documents that would not otherwise have been selected, it also increases the total number of retrieved documents. (We point out at this time that of the three heuristics which we considered, the one which elaborated via the coefficient of association gave the greatest ratio of answer documents to total documents retrieved.) In order to have the advantages of an elaborated request (namely, the relevant

documents that it obtains) and in order to avoid the disadvantages (namely, the larger number of total documents) we now introduce the relevance number to truncate the output lists. That is to say, we use the relevance numbers to separate out the highly relevant from the less relevant documents, by adopting the following rule: Only those documents which are selected by the elaborated request and which have a standardized relevance number greater than 0.5 are to be retrieved. Our experiments with the coefficient of association heuristic show that of a total of 661 documents that were selected by the elaborated requests only 446 (or 67.5 percent) have a standardized relevance number greater than 0.5.[1] Furthermore, if we adopt this rule, then 32 out of the 33 (or 97 percent of the) answer documents which are selected by the automatic elaboration would still be retrieved; i.e., 32 of the 33 answer documents had relevance numbers greater than 0.5.

We conclude by observing that, to a very large degree, the procedures for automatically elaborating upon a request are empirical; i.e., their development and refinement must rest on further empirical testing and experimentation. Hopefully the results of further tests will shed light on and provide new insights into the difficult and exciting problems of information identification, search and retrieval.

[1] For these computations we used a flat a priori probability distribution.